Probability and Statistics

FOURTH EDITION



DEGROOT | SCHERVISH

Probability and Statistics

Fourth Edition

This page intentionally left blank

Probability and Statistics

Fourth Edition

Morris H. DeGroot
Carnegie Mellon University

Mark J. Schervish
Carnegie Mellon University

Addison-Wesley
Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editor in Chief: Deirdre Lynch
Acquisitions Editor: Christopher Cummings
Associate Content Editors: Leah Goldberg, Dana Jones Bettez
Associate Editor: Christina Lepre
Senior Managing Editor: Karen Wernholm
Production Project Manager: Patty Bergin
Cover Designer: Heather Scott
Design Manager: Andrea Nix
Senior Marketing Manager: Alex Gay
Marketing Assistant: Kathleen DeChavez
Senior Author Support/Technology Specialist: Joe Vetere
Rights and Permissions Advisor: Michael Joyce
Manufacturing Manager: Carol Melville
Project Management, Composition: Windfall Software, using ZzTEX
Cover Photo: Shutterstock/© Marilyn Volan

Library of Congress Cataloging-in-Publication Data

DeGroot, Morris H., 1931–1989.
Probability and statistics / Morris H. DeGroot, Mark J. Schervish.—4th ed.
p. cm.
ISBN 978-0-321-50046-5
1. Probabilities—Textbooks. 2. Mathematical statistics—Textbooks.
I. Schervish, Mark J. II. Title.
QA273.D35 2012
519.2—dc22

2010001486

1 2 3 4 5 6 7 8 9 10—EB—14 13 12 11 10

Addison-Wesley
is an imprint of

PEARSON

www.pearsonhighered.com

ISBN 10: 0-321-50046-6
ISBN 13: 978-0-321-50046-5

To the memory of Morrie DeGroot.

MJS

This page intentionally left blank

Contents

Preface

Changes to the Fourth Edition

- I have reorganized many main results that were included in the body of the text by labeling them as theorems in order to facilitate students in finding and referencing these results.
- I have pulled the important defintions and assumptions out of the body of the text and labeled them as such so that they stand out better.
- When a new topic is introduced, I introduce it with a motivating example before delving into the mathematical formalities. Then I return to the example to illustrate the newly introduced material.
- I moved the material on the law of large numbers and the central limit theorem to a new Chapter 6. It seemed more natural to deal with the main large-sample results together.
- I moved the section on Markov chains into Chapter 3. Every time I cover this material with my own students, I stumble over not being able to refer to random variables, distributions, and conditional distributions. I have actually postponed this material until after introducing distributions, and then gone back to cover Markov chains. I feel that the time has come to place it in a more natural location. I also added some material on stationary distributions of Markov chains.
- I have moved the lengthy proofs of several theorems to the ends of their respective sections in order to improve the flow of the presentation of ideas.
- I rewrote Section 7.1 to make the introduction to inference clearer.
- I rewrote Section 9.1 as a more complete introduction to hypothesis testing, including likelihood ratio tests. For instructors not interested in the more mathematical theory of hypothesis testing, it should now be easier to skip from Section 9.1 directly to Section 9.5.

Some other changes that readers will notice:

- I have replaced the notation in which the intersection of two sets A and B had been represented AB with the more popular $A \cap B$. The old notation, although mathematically sound, seemed a bit arcane for a text at this level.
- I added the statements of Stirling's formula and Jensen's inequality.
- I moved the law of total probability and the discussion of partitions of a sample space from Section 2.3 to Section 2.1.
- I define the cumulative distribution function (c.d.f.) as the prefered name of what used to be called only the distribution function (d.f.).
- I added some discussion of histograms in Chapters 3 and 6.
- I rearranged the topics in Sections 3.8 and 3.9 so that simple functions of random variables appear first and the general formulations appear at the end to make it easier for instructors who want to avoid some of the more mathematically challenging parts.
- I emphasized the closeness of a hypergeometric distribution with a large number of available items to a binomial distribution.

- I gave a brief introduction to Chernoff bounds. These are becoming increasingly important in computer science, and their derivation requires only material that is already in the text.
- I changed the definition of confidence interval to refer to the random interval rather than the observed interval. This makes statements less cumbersome, and it corresponds to more modern usage.
- I added a brief discussion of the method of moments in Section 7.6.
- I added brief introductions to Newton's method and the EM algorithm in Chapter 7.
- I introduced the concept of pivotal quantity to facilitate construction of confidence intervals in general.
- I added the statement of the large-sample distribution of the likelihood ratio test statistic. I then used this as an alternative way to test the null hypothesis that two normal means are equal when it is not assumed that the variances are equal.
- I moved the Bonferroni inequality into the main text (Chapter 1) and later (Chapter 11) used it as a way to construct simultaneous tests and confidence intervals.

How to Use This Book

The text is somewhat long for complete coverage in a one-year course at the undergraduate level and is designed so that instructors can make choices about which topics are most important to cover and which can be left for more in-depth study. As an example, many instructors wish to deemphasize the classical counting arguments that are detailed in Sections 1.7–1.9. An instructor who only wants enough information to be able to cover the binomial and/or multinomial distributions can safely discuss only the definitions and theorems on permutations, combinations, and possibly multinomial coefficients. Just make sure that the students realize what these values count, otherwise the associated distributions will make no sense. The various examples in these sections are helpful, but not necessary, for understanding the important distributions. Another example is Section 3.9 on functions of two or more random variables. The use of Jacobians for general multivariate transformations might be more mathematics than the instructors of some undergraduate courses are willing to cover. The entire section could be skipped without causing problems later in the course, but some of the more straightforward cases early in the section (such as convolution) might be worth introducing. The material in Sections 9.2–9.4 on optimal tests in one-parameter families is pretty mathematics, but it is of interest primarily to graduate students who require a very deep understanding of hypothesis testing theory. The rest of Chapter 9 covers everything that an undergraduate course really needs.

In addition to the text, the publisher has an *Instructor's Solutions Manual,* available for download from the Instructor Resource Center at www.pearsonhighered.com/irc, which includes some specific advice about many of the sections of the text. I have taught a year-long probability and statistics sequence from earlier editions of this text for a group of mathematically well-trained juniors and seniors. In the first semester, I covered what was in the earlier edition but is now in the first five chapters (including the material on Markov chains) and parts of Chapter 6. In the second semester, I covered the rest of the new Chapter 6, Chapters 7–9, Sections 11.1–11.5, and Chapter 12. I have also taught a one-semester probability and random processes

course for engineers and computer scientists. I covered what was in the old edition and is now in Chapters 1–6 and 12, including Markov chains, but not Jacobians. This latter course did not emphasize mathematical derivation to the same extent as the course for mathematics students.

A number of sections are designated with an asterisk (*). This indicates that later sections do not rely materially on the material in that section. This designation is not intended to suggest that instructors skip these sections. Skipping one of these sections will not cause the students to miss definitions or results that they will need later. The sections are 2.4, 3.10, 4.8, 7.7, 7.8, 7.9, 8.6, 8.8, 9.2, 9.3, 9.4, 9.8, 9.9, 10.6, 10.7, 10.8, 11.4, 11.7, 11.8, and 12.5. Aside from cross-references between sections within this list, occasional material from elsewhere in the text does refer back to some of the sections in this list. Each of the dependencies is quite minor, however. Most of the dependencies involve references from Chapter 12 back to one of the optional sections. The reason for this is that the optional sections address some of the more difficult material, and simulation is most useful for solving those difficult problems that cannot be solved analytically. Except for passing references that help put material into context, the dependencies are as follows:

- The sample distribution function (Section 10.6) is reintroduced during the discussion of the bootstrap in Section 12.6. The sample distribution function is also a useful tool for displaying simulation results. It could be introduced as early as Example 12.3.7 simply by covering the first subsection of Section 10.6.
- The material on robust estimation (Section 10.7) is revisited in some simulation exercises in Section 12.2 (Exercises 4, 5, 7, and 8).
- Example 12.3.4 makes reference to the material on two-way analysis of variance (Sections 11.7 and 11.8).

Supplements

The text is accompanied by the following supplementary material:

- **Instructor's Solutions Manual** contains fully worked solutions to all exercises in the text. Available for download from the Instructor Resource Center at www.pearsonhighered.com/irc.
- **Student Solutions Manual** contains fully worked solutions to all odd exercises in the text. Available for purchase from MyPearsonStore at www.mypearsonstore .com. (ISBN-13: 978-0-321-71598-2; ISBN-10: 0-321-71598-5)

Acknowledgments

There are many people that I want to thank for their help and encouragement during this revision. First and foremost, I want to thank Marilyn DeGroot and Morrie's children for giving me the chance to revise Morrie's masterpiece.

I am indebted to the many readers, reviewers, colleagues, staff, and people at Addison-Wesley whose help and comments have strengthened this edition. The reviewers were:

Andre Adler, Illinois Institute of Technology; E. N. Barron, Loyola University; Brian Blank, Washington University in St. Louis; Indranil Chakraborty, University of Oklahoma; Daniel Chambers, Boston College; Rita Chattopadhyay, Eastern Michigan University; Stephen A. Chiappari, Santa Clara University; Sheng-Kai Chang, Wayne State University; Justin Corvino, Lafayette College; Michael Evans, University of

Toronto; Doug Frank, Indiana University of Pennsylvania; Anda Gadidov, Kennesaw State University; Lyn Geisler, Randolph–Macon College; Prem Goel, Ohio State University; Susan Herring, Sonoma State University; Pawel Hitczenko, Drexel University; Lifang Hsu, Le Moyne College; Wei-Min Huang, Lehigh University; Syed Kirmani, University of Northern Iowa; Michael Lavine, Duke University; Rich Levine, San Diego State University; John Liukkonen, Tulane University; Sergio Loch, Grand View College; Rosa Matzkin, Northwestern University; Terry McConnell, Syracuse University; Hans-Georg Mueller, University of California–Davis; Robert Myers, Bethel College; Mario Peruggia, The Ohio State University; Stefan Ralescu, Queens University; Krishnamurthi Ravishankar, SUNY New Paltz; Diane Saphire, Trinity University; Steven Sepanski, Saginaw Valley State University; Hen-Siong Tan, Pennsylvania University; Kanapathi Thiru, University of Alaska; Kenneth Troske, Johns Hopkins University; John Van Ness, University of Texas at Dallas; Yehuda Vardi, Rutgers University; Yelena Vaynberg, Wayne State University; Joseph Verducci, Ohio State University; Mahbobeh Vezveai, Kent State University; Brani Vidakovic, Duke University; Karin Vorwerk, Westfield State College; Bette Warren, Eastern Michigan University; Calvin L. Williams, Clemson University; Lori Wolff, University of Mississippi.

The person who checked the accuracy of the book was Anda Gadidov, Kennesaw State University. I would also like to thank my colleagues at Carnegie Mellon University, especially Anthony Brockwell, Joel Greenhouse, John Lehoczky, Heidi Sestrich, and Valerie Ventura.

The people at Addison-Wesley and other organizations that helped produce the book were Paul Anagnostopoulos, Patty Bergin, Dana Jones Bettez, Chris Cummings, Kathleen DeChavez, Alex Gay, Leah Goldberg, Karen Hartpence, and Christina Lepre.

If I left anyone out, it was unintentional, and I apologize. Errors inevitably arise in any project like this (meaning a project in which I am involved). For this reason, I shall post information about the book, including a list of corrections, on my Web page, http://www.stat.cmu.edu/~mark/, as soon as the book is published. Readers are encouraged to send me any errors that they discover.

Mark J. Schervish
October 2010

Chapter 1

Introduction to Probability

1.1 The History of Probability

The use of probability to measure uncertainty and variability dates back hundreds of years. Probability has found application in areas as diverse as medicine, gambling, weather forecasting, and the law.

The concepts of chance and uncertainty are as old as civilization itself. People have always had to cope with uncertainty about the weather, their food supply, and other aspects of their environment, and have striven to reduce this uncertainty and its effects. Even the idea of gambling has a long history. By about the year 3500 B.C., games of chance played with bone objects that could be considered precursors of dice were apparently highly developed in Egypt and elsewhere. Cubical dice with markings virtually identical to those on modern dice have been found in Egyptian tombs dating from 2000 B.C. We know that gambling with dice has been popular ever since that time and played an important part in the early development of probability theory.

It is generally believed that the mathematical theory of probability was started by the French mathematicians Blaise Pascal (1623–1662) and Pierre Fermat (1601–1665) when they succeeded in deriving exact probabilities for certain gambling problems involving dice. Some of the problems that they solved had been outstanding for about 300 years. However, numerical probabilities of various dice combinations had been calculated previously by Girolamo Cardano (1501–1576) and Galileo Galilei (1564–1642).

The theory of probability has been developed steadily since the seventeenth century and has been widely applied in diverse fields of study. Today, probability theory is an important tool in most areas of engineering, science, and management. Many research workers are actively engaged in the discovery and establishment of new applications of probability in fields such as medicine, meteorology, photography from satellites, marketing, earthquake prediction, human behavior, the design of computer systems, finance, genetics, and law. In many legal proceedings involving antitrust violations or employment discrimination, both sides will present probability and statistical calculations to help support their cases.

References

The ancient history of gambling and the origins of the mathematical theory of probability are discussed by David (1988), Ore (1960), Stigler (1986), and Todhunter (1865).

Some introductory books on probability theory, which discuss many of the same topics that will be studied in this book, are Feller (1968); Hoel, Port, and Stone (1971); Meyer (1970); and Olkin, Gleser, and Derman (1980). Other introductory books, which discuss both probability theory and statistics at about the same level as they will be discussed in this book, are Brunk (1975); Devore (1999); Fraser (1976); Hogg and Tanis (1997); Kempthorne and Folks (1971); Larsen and Marx (2001); Larson (1974); Lindgren (1976); Miller and Miller (1999); Mood, Graybill, and Boes (1974); Rice (1995); and Wackerly, Mendenhall, and Schaeffer (2008).

1.2 Interpretations of Probability

This section describes three common operational interpretations of probability. Although the interpretations may seem incompatible, it is fortunate that the calculus of probability (the subject matter of the first six chapters of this book) applies equally well no matter which interpretation one prefers.

In addition to the many formal applications of probability theory, the concept of probability enters our everyday life and conversation. We often hear and use such expressions as "It probably will rain tomorrow afternoon," "It is very likely that the plane will arrive late," or "The chances are good that he will be able to join us for dinner this evening." Each of these expressions is based on the concept of the probability, or the likelihood, that some specific event will occur.

Despite the fact that the concept of probability is such a common and natural part of our experience, no single scientific interpretation of the term *probability* is accepted by all statisticians, philosophers, and other authorities. Through the years, each interpretation of probability that has been proposed by some authorities has been criticized by others. Indeed, the true meaning of probability is still a highly controversial subject and is involved in many current philosophical discussions pertaining to the foundations of statistics. Three different interpretations of probability will be described here. Each of these interpretations can be very useful in applying probability theory to practical problems.

The Frequency Interpretation of Probability

In many problems, the probability that some specific outcome of a process will be obtained can be interpreted to mean the *relative frequency* with which that outcome would be obtained if the process were repeated a large number of times under similar conditions. For example, the probability of obtaining a head when a coin is tossed is considered to be 1/2 because the relative frequency of heads should be approximately 1/2 when the coin is tossed a large number of times under similar conditions. In other words, it is assumed that the proportion of tosses on which a head is obtained would be approximately 1/2.

Of course, the conditions mentioned in this example are too vague to serve as the basis for a scientific definition of probability. First, a "large number" of tosses of the coin is specified, but there is no definite indication of an actual number that would

be considered large enough. Second, it is stated that the coin should be tossed each time "under similar conditions," but these conditions are not described precisely. The conditions under which the coin is tossed must not be completely identical for each toss because the outcomes would then be the same, and there would be either all heads or all tails. In fact, a skilled person can toss a coin into the air repeatedly and catch it in such a way that a head is obtained on almost every toss. Hence, the tosses must not be completely controlled but must have some "random" features.

Furthermore, it is stated that the relative frequency of heads should be "approximately 1/2," but no limit is specified for the permissible variation from 1/2. If a coin were tossed 1,000,000 times, we would not expect to obtain exactly 500,000 heads. Indeed, we would be extremely surprised if we obtained exactly 500,000 heads. On the other hand, neither would we expect the number of heads to be very far from 500,000. It would be desirable to be able to make a precise statement of the likelihoods of the different possible numbers of heads, but these likelihoods would of necessity depend on the very concept of probability that we are trying to define.

Another shortcoming of the frequency interpretation of probability is that it applies only to a problem in which there can be, at least in principle, a large number of similar repetitions of a certain process. Many important problems are not of this type. For example, the frequency interpretation of probability cannot be applied directly to the probability that a specific acquaintance will get married within the next two years or to the probability that a particular medical research project will lead to the development of a new treatment for a certain disease within a specified period of time.

The Classical Interpretation of Probability

The classical interpretation of probability is based on the concept of *equally likely outcomes*. For example, when a coin is tossed, there are two possible outcomes: a head or a tail. If it may be assumed that these outcomes are equally likely to occur, then they must have the same probability. Since the sum of the probabilities must be 1, both the probability of a head and the probability of a tail must be 1/2. More generally, if the outcome of some process must be one of n different outcomes, and if these n outcomes are equally likely to occur, then the probability of each outcome is $1/n$.

Two basic difficulties arise when an attempt is made to develop a formal definition of probability from the classical interpretation. First, the concept of equally likely outcomes is essentially based on the concept of probability that we are trying to define. The statement that two possible outcomes are equally likely to occur is the same as the statement that two outcomes have the same probability. Second, no systematic method is given for assigning probabilities to outcomes that are not assumed to be equally likely. When a coin is tossed, or a well-balanced die is rolled, or a card is chosen from a well-shuffled deck of cards, the different possible outcomes can usually be regarded as equally likely because of the nature of the process. However, when the problem is to guess whether an acquaintance will get married or whether a research project will be successful, the possible outcomes would not typically be considered to be equally likely, and a different method is needed for assigning probabilities to these outcomes.

The Subjective Interpretation of Probability

According to the subjective, or personal, interpretation of probability, the probability that a person assigns to a possible outcome of some process represents her own

judgment of the likelihood that the outcome will be obtained. This judgment will be based on each person's beliefs and information about the process. Another person, who may have different beliefs or different information, may assign a different probability to the same outcome. For this reason, it is appropriate to speak of a certain person's *subjective probability* of an outcome, rather than to speak of the *true probability* of that outcome.

As an illustration of this interpretation, suppose that a coin is to be tossed once. A person with no special information about the coin or the way in which it is tossed might regard a head and a tail to be equally likely outcomes. That person would then assign a subjective probability of 1/2 to the possibility of obtaining a head. The person who is actually tossing the coin, however, might feel that a head is much more likely to be obtained than a tail. In order that people in general may be able to assign subjective probabilities to the outcomes, they must express the strength of their belief in numerical terms. Suppose, for example, that they regard the likelihood of obtaining a head to be the same as the likelihood of obtaining a red card when one card is chosen from a well-shuffled deck containing four red cards and one black card. Because those people would assign a probability of 4/5 to the possibility of obtaining a red card, they should also assign a probability of 4/5 to the possibility of obtaining a head when the coin is tossed.

This subjective interpretation of probability can be formalized. In general, if people's judgments of the relative likelihoods of various combinations of outcomes satisfy certain conditions of consistency, then it can be shown that their subjective probabilities of the different possible events can be uniquely determined. However, there are two difficulties with the subjective interpretation. First, the requirement that a person's judgments of the relative likelihoods of an infinite number of events be completely consistent and free from contradictions does not seem to be humanly attainable, unless a person is simply willing to adopt a collection of judgments known to be consistent. Second, the subjective interpretation provides no "objective" basis for two or more scientists working together to reach a common evaluation of the state of knowledge in some scientific area of common interest.

On the other hand, recognition of the subjective interpretation of probability has the salutary effect of emphasizing some of the subjective aspects of science. A particular scientist's evaluation of the probability of some uncertain outcome must ultimately be that person's own evaluation based on all the evidence available. This evaluation may well be based in part on the frequency interpretation of probability, since the scientist may take into account the relative frequency of occurrence of this outcome or similar outcomes in the past. It may also be based in part on the classical interpretation of probability, since the scientist may take into account the total number of possible outcomes that are considered equally likely to occur. Nevertheless, the final assignment of numerical probabilities is the responsibility of the scientist herself.

The subjective nature of science is also revealed in the actual problem that a particular scientist chooses to study from the class of problems that might have been chosen, in the experiments that are selected in carrying out this study, and in the conclusions drawn from the experimental data. The mathematical theory of probability and statistics can play an important part in these choices, decisions, and conclusions.

Note: The Theory of Probability Does Not Depend on Interpretation. The mathematical theory of probability is developed and presented in Chapters 1–6 of this book without regard to the controversy surrounding the different interpretations of

the term probability. This theory is correct and can be usefully applied, regardless of which interpretation of probability is used in a particular problem. The theories and techniques that will be presented in this book have served as valuable guides and tools in almost all aspects of the design and analysis of effective experimentation.

1.3 Experiments and Events

Probability will be the way that we quantify how likely something is to occur (in the sense of one of the interpretations in Sec. 1.2). In this section, we give examples of the types of situations in which probability will be used.

Types of Experiments

The theory of probability pertains to the various possible outcomes that might be obtained and the possible events that might occur when an experiment is performed.

Definition 1.3.1 Experiment and Event. An *experiment* is any process, real or hypothetical, in which the possible outcomes can be identified ahead of time. An *event* is a well-defined set of possible outcomes of the experiment.

The breadth of this definition allows us to call almost any imaginable process an experiment whether or not its outcome will ever be known. The probability of each event will be our way of saying how likely it is that the outcome of the experiment is in the event. Not every set of possible outcomes will be called an event. We shall be more specific about which subsets count as events in Sec. 1.4.

Probability will be most useful when applied to a real experiment in which the outcome is not known in advance, but there are many hypothetical experiments that provide useful tools for modeling real experiments. A common type of hypothetical experiment is repeating a well-defined task infinitely often under similar conditions. Some examples of experiments and specific events are given next. In each example, the words following "the probability that" describe the event of interest.

1. In an experiment in which a coin is to be tossed 10 times, the experimenter might want to determine the probability that at least four heads will be obtained.
2. In an experiment in which a sample of 1000 transistors is to be selected from a large shipment of similar items and each selected item is to be inspected, a person might want to determine the probability that not more than one of the selected transistors will be defective.
3. In an experiment in which the air temperature at a certain location is to be observed every day at noon for 90 successive days, a person might want to determine the probability that the average temperature during this period will be less than some specified value.
4. From information relating to the life of Thomas Jefferson, a person might want to determine the probability that Jefferson was born in the year 1741.
5. In evaluating an industrial research and development project at a certain time, a person might want to determine the probability that the project will result in the successful development of a new product within a specified number of months.

The Mathematical Theory of Probability

As was explained in Sec. 1.2, there is controversy in regard to the proper meaning and interpretation of some of the probabilities that are assigned to the outcomes of many experiments. However, once probabilities have been assigned to some simple outcomes in an experiment, there is complete agreement among all authorities that the mathematical theory of probability provides the appropriate methodology for the further study of these probabilities. Almost all work in the mathematical theory of probability, from the most elementary textbooks to the most advanced research, has been related to the following two problems: (i) methods for determining the probabilities of certain events from the specified probabilities of each possible outcome of an experiment and (ii) methods for revising the probabilities of events when additional relevant information is obtained.

These methods are based on standard mathematical techniques. The purpose of the first six chapters of this book is to present these techniques, which, together, form the mathematical theory of probability.

1.4 Set Theory

This section develops the formal mathematical model for events, namely, the theory of sets. Several important concepts are introduced, namely, element, subset, empty set, intersection, union, complement, and disjoint sets.

The Sample Space

Definition 1.4.1 Sample Space. The collection of all possible outcomes of an experiment is called the *sample space* of the experiment.

The sample space of an experiment can be thought of as a *set*, or collection, of different possible outcomes; and each outcome can be thought of as a *point*, or an *element*, in the sample space. Similarly, events can be thought of as *subsets* of the sample space.

Example 1.4.1 Rolling a Die. When a six-sided die is rolled, the sample space can be regarded as containing the six numbers 1, 2, 3, 4, 5, 6, each representing a possible side of the die that shows after the roll. Symbolically, we write

$$S = \{1, 2, 3, 4, 5, 6\}.$$

One event A is that an even number is obtained, and it can be represented as the subset $A = \{2, 4, 6\}$. The event B that a number greater than 2 is obtained is defined by the subset $B = \{3, 4, 5, 6\}$. ◀

Because we can interpret outcomes as elements of a set and events as subsets of a set, the language and concepts of set theory provide a natural context for the development of probability theory. The basic ideas and notation of set theory will now be reviewed.

Relations of Set Theory

Let S denote the sample space of some experiment. Then each possible outcome s of the experiment is said to be a member of the space S, or to belong to the space S. The statement that s is a member of S is denoted symbolically by the relation $s \in S$.

When an experiment has been performed and we say that some event E has occurred, we mean two equivalent things. One is that the outcome of the experiment satisfied the conditions that specified that event E. The other is that the outcome, considered as a point in the sample space, is an element of E.

To be precise, we should say which sets of outcomes correspond to events as defined above. In many applications, such as Example 1.4.1, it will be clear which sets of outcomes should correspond to events. In other applications (such as Example 1.4.5 coming up later), there are too many sets available to have them all be events. Ideally, we would like to have the largest possible collection of sets called events so that we have the broadest possible applicability of our probability calculations. However, when the sample space is too large (as in Example 1.4.5) the theory of probability simply will not extend to the collection of all subsets of the sample space. We would prefer not to dwell on this point for two reasons. First, a careful handling requires mathematical details that interfere with an initial understanding of the important concepts, and second, the practical implications for the results in this text are minimal. In order to be mathematically correct without imposing an undue burden on the reader, we note the following. In order to be able to do all of the probability calculations that we might find interesting, there are three simple conditions that must be met by the collection of sets that we call events. In every problem that we see in this text, there exists a collection of sets that includes all the sets that we will need to discuss and that satisfies the three conditions, and the reader should assume that such a collection has been chosen as the events. For a sample space S with only finitely many outcomes, the collection of all subsets of S satisfies the conditions, as the reader can show in Exercise 12 in this section.

The first of the three conditions can be stated immediately.

Condition 1 The sample space S must be an event.

That is, we must include the sample space S in our collection of events. The other two conditions will appear later in this section because they require additional definitions. Condition 2 is on page 9, and Condition 3 is on page 10.

Definition 1.4.2 Containment. It is said that a set A *is contained in* another set B if every element of the set A also belongs to the set B. This relation between two events is expressed symbolically by the expression $A \subset B$, which is the set-theoretic expression for saying that A is a subset of B. Equivalently, if $A \subset B$, we may say that B *contains* A and may write $B \supset A$.

For events, to say that $A \subset B$ means that if A occurs then so does B.

The proof of the following result is straightforward and is omitted.

Theorem 1.4.1 Let A, B, and C be events. Then $A \subset S$. If $A \subset B$ and $B \subset A$, then $A = B$. If $A \subset B$ and $B \subset C$, then $A \subset C$. ■

Example 1.4.2 Rolling a Die. In Example 1.4.1, suppose that A is the event that an even number is obtained and C is the event that a number greater than 1 is obtained. Since $A = \{2, 4, 6\}$ and $C = \{2, 3, 4, 5, 6\}$, it follows that $A \subset C$. ◀

The Empty Set Some events are impossible. For example, when a die is rolled, it is impossible to obtain a negative number. Hence, the event that a negative number will be obtained is defined by the subset of S that contains no outcomes.

Definition 1.4.3 Empty Set. The subset of S that contains no elements is called the *empty set*, or *null set*, and it is denoted by the symbol $\emptyset$.

In terms of events, the empty set is any event that cannot occur.

Theorem 1.4.2 Let A be an event. Then $\emptyset \subset A$.

Proof Let A be an arbitrary event. Since the empty set $\emptyset$ contains no points, it is logically correct to say that every point belonging to $\emptyset$ also belongs to A, or $\emptyset \subset A$. ■

Finite and Infinite Sets Some sets contain only finitely many elements, while others have infinitely many elements. There are two sizes of infinite sets that we need to distinguish.

Definition 1.4.4 Countable/Uncountable. An infinite set A is *countable* if there is a one-to-one correspondence between the elements of A and the set of natural numbers $\{1, 2, 3, \ldots\}$. A set is *uncountable* if it is neither finite nor countable. If we say that a set has *at most countably many* elements, we mean that the set is either finite or countable.

Examples of countably infinite sets include the integers, the even integers, the odd integers, the prime numbers, and any infinite sequence. Each of these can be put in one-to-one correspondence with the natural numbers. For example, the following function f puts the integers in one-to-one correspondence with the natural numbers:

$$f(n) = \begin{cases} \frac{n-1}{2} & \text{if } n \text{ is odd,} \\ -\frac{n}{2} & \text{if } n \text{ is even.} \end{cases}$$

Every infinite sequence of distinct items is a countable set, as its indexing puts it in one-to-one correspondence with the natural numbers. Examples of uncountable sets include the real numbers, the positive reals, the numbers in the interval $[0, 1]$, and the set of all ordered pairs of real numbers. An argument to show that the real numbers are uncountable appears at the end of this section. Every subset of the integers has at most countably many elements.

Operations of Set Theory

Definition 1.4.5 Complement. The *complement* of a set A is defined to be the set that contains all elements of the sample space S that *do not* belong to A. The notation for the complement of A is A^c.

In terms of events, the event A^c is the event that A does not occur.

Example 1.4.3 Rolling a Die. In Example 1.4.1, suppose again that A is the event that an even number is rolled; then $A^c = \{1, 3, 5\}$ is the event that an odd number is rolled. ◄

We can now state the second condition that we require of the collection of events.

Figure 1.1 The event A^c.

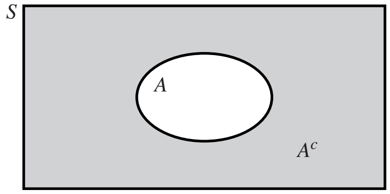
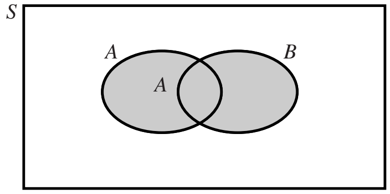


Figure 1.2 The set $A \cup B$.



Condition 2 If A is an event, then A^c is also an event.

That is, for each set A of outcomes that we call an event, we must also call its complement A^c an event.

A generic version of the relationship between A and A^c is sketched in Fig. 1.1. A sketch of this type is called a *Venn diagram*.

Some properties of the complement are stated without proof in the next result.

Theorem 1.4.3 Let A be an event. Then

$$(A^c)^c = A, \qquad \emptyset^c = S, \qquad S^c = \emptyset.$$

The empty event $\emptyset$ is an event. ■

Definition 1.4.6 Union of Two Sets. If A and B are any two sets, the *union* of A and B is defined to be the set containing all outcomes that belong to A alone, to B alone, or to both A and B. The notation for the union of A and B is $A \cup B$.

The set $A \cup B$ is sketched in Fig. 1.2. In terms of events, $A \cup B$ is the event that either A or B or both occur.

The union has the following properties whose proofs are left to the reader.

Theorem 1.4.4 For all sets A and B,

$$A \cup B = B \cup A, \qquad A \cup A = A, \qquad A \cup A^c = S,$$
$$A \cup \emptyset = A, \qquad A \cup S = S.$$

Furthermore, if $A \subset B$, then $A \cup B = B$. ■

The concept of union extends to more than two sets.

Definition 1.4.7 Union of Many Sets. The *union* of n sets $A_1, \ldots, A_n$ is defined to be the set that contains all outcomes that belong to at least one of these n sets. The notation for this union is either of the following:

$$A_1 \cup A_2 \cup \cdots \cup A_n \text{ or } \bigcup_{i=1}^{n} A_i.$$

Similarly, the *union* of an infinite sequence of sets $A_1, A_2, \ldots$ is the set that contains all outcomes that belong to at least one of the events in the sequence. The infinite union is denoted by $\bigcup_{i=1}^{\infty} A_i$.

In terms of events, the union of a collection of events is the event that at least one of the events in the collection occurs.

We can now state the final condition that we require for the collection of sets that we call events.

Condition 3 If $A_1, A_2, \ldots$ is a countable collection of events, then $\bigcup_{i=1}^{\infty} A_i$ is also an event.

In other words, if we choose to call each set of outcomes in some countable collection an event, we are required to call their union an event also. We do *not* require that the union of an arbitrary collection of events be an event. To be clear, let I be an arbitrary set that we use to index a general collection of events $\{A_i : i \in I\}$. The union of the events in this collection is the set of outcomes that are in at least one of the events in the collection. The notation for this union is $\bigcup_{i \in I} A_i$. We do not require that $\bigcup_{i \in I} A_i$ be an event unless I is countable.

Condition 3 refers to a countable collection of events. We can prove that the condition also applies to every finite collection of events.

Theorem 1.4.5 The union of a finite number of events $A_1, \ldots, A_n$ is an event.

Proof For each $m = n + 1, n + 2, \ldots$, define $A_m = \emptyset$. Because $\emptyset$ is an event, we now have a countable collection $A_1, A_2, \ldots$ of events. It follows from Condition 3 that $\bigcup_{m=1}^{\infty} A_m$ is an event. But it is easy to see that $\bigcup_{m=1}^{\infty} A_m = \bigcup_{m=1}^{n} A_m$. ■

The union of three events A, B, and C can be constructed either directly from the definition of $A \cup B \cup C$ or by first evaluating the union of any two of the events and then forming the union of this combination of events and the third event. In other words, the following result is true.

Theorem 1.4.6 Associative Property. For every three events A, B, and C, the following associative relations are satisfied:

$$A \cup B \cup C = (A \cup B) \cup C = A \cup (B \cup C).$$ ■

Definition 1.4.8 Intersection of Two Sets. If A and B are any two sets, the *intersection* of A and B is defined to be the set that contains all outcomes that belong *both to A and to B*. The notation for the intersection of A and B is $A \cap B$.

The set $A \cap B$ is sketched in a Venn diagram in Fig. 1.3. In terms of events, $A \cap B$ is the event that both A and B occur.

The proof of the first part of the next result follows from Exercise 3 in this section. The rest of the proof is straightforward.

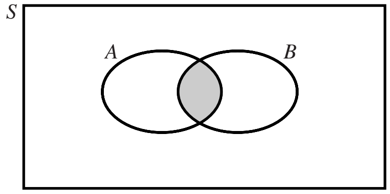


Figure 1.3 The set $A \cap B$.

Theorem 1.4.7 If A and B are events, then so is $A \cap B$. For all events A and B,

$$A \cap B = B \cap A, \qquad A \cap A = A, \qquad A \cap A^c = \emptyset,$$
$$A \cap \emptyset = \emptyset, \qquad A \cap S = A.$$

Furthermore, if $A \subset B$, then $A \cap B = A$. ■

The concept of intersection extends to more than two sets.

Definition 1.4.9 Intersection of Many Sets. The *intersection* of n sets $A_1, \ldots, A_n$ is defined to be the set that contains the elements that are common to all these n sets. The notation for this intersection is $A_1 \cap A_2 \cap \ldots \cap A_n$ or $\bigcap_{i=1}^{n} A_i$. Similar notations are used for the intersection of an infinite sequence of sets or for the intersection of an arbitrary collection of sets.

In terms of events, the intersection of a collection of events is the event that every event in the collection occurs.

The following result concerning the intersection of three events is straightforward to prove.

Theorem 1.4.8 Associative Property. For every three events A, B, and C, the following associative relations are satisfied:

$$A \cap B \cap C = (A \cap B) \cap C = A \cap (B \cap C).$$ ■

Definition 1.4.10 Disjoint/Mutually Exclusive. It is said that two sets A and B are *disjoint*, or *mutually exclusive*, if A and B have no outcomes in common, that is, if $A \cap B = \emptyset$. The sets $A_1, \ldots, A_n$ or the sets $A_1, A_2, \ldots$ are disjoint if for every $i \neq j$, we have that A_i and A_j are disjoint, that is, $A_i \cap A_j = \emptyset$ for all $i \neq j$. The events in an arbitrary collection are disjoint if no two events in the collection have any outcomes in common.

In terms of events, A and B are disjoint if they cannot both occur.

As an illustration of these concepts, a Venn diagram for three events A_1, A_2, and A_3 is presented in Fig. 1.4. This diagram indicates that the various intersections of A_1, A_2, and A_3 and their complements will partition the sample space S into eight disjoint subsets.

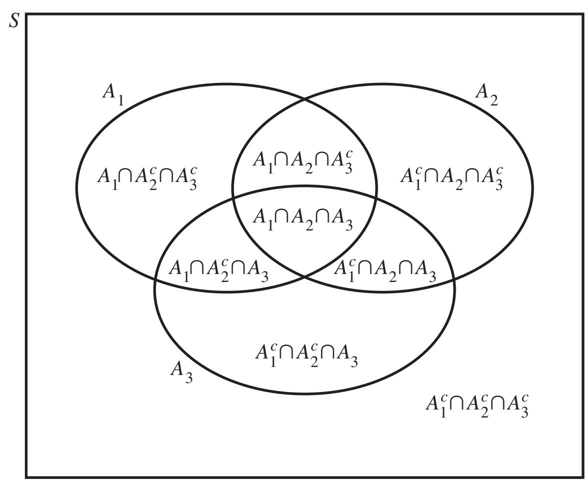


Figure 1.4 Partition of S determined by three events A_1, A_2, A_3.

Example 1.4.4

Tossing a Coin. Suppose that a coin is tossed three times. Then the sample space S contains the following eight possible outcomes $s_1, \ldots, s_8$:

s_1: HHH,
s_2: THH,
s_3: HTH,
s_4: HHT,
s_5: HTT,
s_6: THT,
s_7: TTH,
s_8: TTT.

In this notation, H indicates a head and T indicates a tail. The outcome s_3, for example, is the outcome in which a head is obtained on the first toss, a tail is obtained on the second toss, and a head is obtained on the third toss.

To apply the concepts introduced in this section, we shall define four events as follows: Let A be the event that at least one head is obtained in the three tosses; let B be the event that a head is obtained on the second toss; let C be the event that a tail is obtained on the third toss; and let D be the event that *no* heads are obtained. Accordingly,

$$\begin{aligned} A &= \{s_1, s_2, s_3, s_4, s_5, s_6, s_7\}, \\ B &= \{s_1, s_2, s_4, s_6\}, \\ C &= \{s_4, s_5, s_6, s_8\}, \\ D &= \{s_8\}. \end{aligned}$$

Various relations among these events can be derived. Some of these relations are $B \subset A$, $A^c = D$, $B \cap D = \emptyset$, $A \cup C = S$, $B \cap C = \{s_4, s_6\}$, $(B \cup C)^c = \{s_3, s_7\}$, and $A \cap (B \cup C) = \{s_1, s_2, s_4, s_5, s_6\}$. ◀

Example 1.4.5

Demands for Utilities. A contractor is building an office complex and needs to plan for water and electricity demand (sizes of pipes, conduit, and wires). After consulting with prospective tenants and examining historical data, the contractor decides that the demand for electricity will range somewhere between 1 million and 150 million kilowatt-hours per day and water demand will be between 4 and 200 (in thousands of gallons per day). All combinations of electrical and water demand are considered possible. The shaded region in Fig. 1.5 shows the sample space for the experiment, consisting of learning the actual water and electricity demands for the office complex. We can express the sample space as the set of ordered pairs $\{(x, y) : 4 \le x \le 200, 1 \le y \le 150\}$, where x stands for water demand in thousands of gallons per day and y

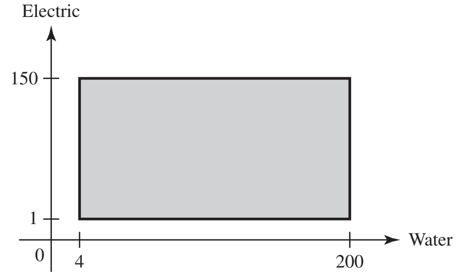


Figure 1.5 Sample space for water and electric demand in Example 1.4.5

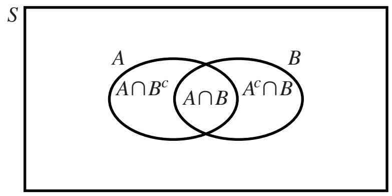


Figure 1.6 Partition of $A \cup B$ in Theorem 1.4.11.

stands for the electric demand in millions of kilowatt-hours per day. The types of sets that we want to call events include sets like

$$\{\text{water demand is at least } 100\} = \{(x, y) : x \geq 100\}, \text{ and}$$
$$\{\text{electric demand is no more than } 35\} = \{(x, y) : y \leq 35\},$$

along with intersections, unions, and complements of such sets. This sample space has infinitely many points. Indeed, the sample space is uncountable. There are many more sets that are difficult to describe and which we will have no need to consider as events. ◀

Additional Properties of Sets The proof of the following useful result is left to Exercise 3 in this section.

Theorem 1.4.9 De Morgan's Laws. For every two sets A and B,

$$(A \cup B)^c = A^c \cap B^c \quad \text{and} \quad (A \cap B)^c = A^c \cup B^c. \quad \blacksquare$$

The generalization of Theorem 1.4.9 is the subject of Exercise 5 in this section.

The proofs of the following distributive properties are left to Exercise 2 in this section. These properties also extend in natural ways to larger collections of events.

Theorem 1.4.10 Distributive Properties. For every three sets A, B, and C,

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad \text{and} \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C). \quad \blacksquare$$

The following result is useful for computing probabilities of events that can be partitioned into smaller pieces. Its proof is left to Exercise 4 in this section, and is illuminated by Fig. 1.6.

Theorem 1.4.11 Partitioning a Set. For every two sets A and B, $A \cap B$ and $A \cap B^c$ are disjoint and

$$A = (A \cap B) \cup (A \cap B^c).$$

In addition, B and $A \cap B^c$ are disjoint, and

$$A \cup B = B \cup (A \cap B^c). \quad \blacksquare$$

Proof That the Real Numbers Are Uncountable

We shall show that the real numbers in the interval $[0, 1)$ are uncountable. Every larger set is a fortiori uncountable. For each number $x \in [0, 1)$, define the sequence $\{a_n(x)\}_{n=1}^{\infty}$ as follows. First, $a_1(x) = \lfloor 10x \rfloor$, where $\lfloor y \rfloor$ stands for the greatest integer less than or equal to y (round nonintegers down to the closest integer below). Then

$$\begin{array}{cccccccc}
\underline{0} & 2 & 3 & 0 & 7 & 1 & 3 & \ldots \\
1 & \underline{9} & 9 & 2 & 1 & 0 & 0 & \ldots \\
2 & 7 & \underline{3} & 6 & 0 & 1 & 1 & \ldots \\
8 & 0 & 2 & \underline{1} & 2 & 7 & 9 & \ldots \\
7 & 0 & 1 & 6 & \underline{0} & 1 & 3 & \ldots \\
1 & 5 & 1 & 5 & \underline{1} & 5 & 1 & \ldots \\
2 & 3 & 4 & 5 & 6 & \underline{7} & 8 & \ldots \\
0 & 1 & 7 & 3 & 2 & 9 & \underline{8} & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

Figure 1.7 An array of a countable collection of sequences of digits with the diagonal underlined.

set $b_1(x) = 10x - a_1(x)$, which will again be in $[0, 1)$. For $n > 1$, $a_n(x) = \lfloor 10b_{n-1}(x) \rfloor$ and $b_n(x) = 10b_{n-1}(x) - a_n(x)$. It is easy to see that the sequence $\{a_n(x)\}_{n=1}^{\infty}$ gives a decimal expansion for x in the form

$$x = \sum_{n=1}^{\infty} a_n(x)10^{-n}. \tag{1.4.1}$$

By construction, each number of the form $x = k/10^m$ for some nonnegative integers k and m will have $a_n(x) = 0$ for $n > m$. The numbers of the form $k/10^m$ are the only ones that have an alternate decimal expansion $x = \sum_{n=1}^{\infty} c_n(x)10^{-n}$. When k is not a multiple of 10, this alternate expansion satisfies $c_n(x) = a_n(x)$ for $n = 1, \ldots, m-1$, $c_m(x) = a_m(x) - 1$, and $c_n(x) = 9$ for $n > m$. Let $C = \{0, 1, \ldots, 9\}^{\infty}$ stand for the set of all infinite sequences of digits. Let B denote the subset of C consisting of those sequences that don't end in repeating 9's. Then we have just constructed a function a from the interval $[0, 1)$ onto B that is one-to-one and whose inverse is given in (1.4.1). We now show that the set B is uncountable, hence $[0, 1)$ is uncountable. Take any countable subset of B and arrange the sequences into a rectangular array with the kth sequence running across the kth row of the array for $k = 1, 2, \ldots$. Figure 1.7 gives an example of part of such an array.

In Fig. 1.7, we have underlined the kth digit in the kth sequence for each k. This portion of the array is called the *diagonal* of the array. We now show that there must exist a sequence in B that is not part of this array. This will prove that the whole set B cannot be put into such an array, and hence cannot be countable. Construct the sequence $\{d_n\}_{n=1}^{\infty}$ as follows. For each n, let $d_n = 2$ if the nth digit in the nth sequence is 1, and $d_n = 1$ otherwise. This sequence does not end in repeating 9's; hence, it is in B. We conclude the proof by showing that $\{d_n\}_{n=1}^{\infty}$ does not appear anywhere in the array. If the sequence did appear in the array, say, in the kth row, then its kth element would be the kth diagonal element of the array. But we constructed the sequence so that for every n (including $n = k$), its nth element never matched the nth diagonal element. Hence, the sequence can't be in the kth row, no matter what k is. The argument given here is essentially that of the nineteenth-century German mathematician Georg Cantor.

Summary

We will use set theory for the mathematical model of events. Outcomes of an experiment are elements of some sample space S, and each event is a subset of S. Two events both occur if the outcome is in the intersection of the two sets. At least one of a collection of events occurs if the outcome is in the union of the sets. Two events cannot both occur if the sets are disjoint. An event fails to occur if the outcome is in the complement of the set. The empty set stands for every event that cannot possibly occur. The collection of events is assumed to contain the sample space, the complement of each event, and the union of each countable collection of events.

Exercises

1. Suppose that $A \subset B$. Show that $B^c \subset A^c$.

2. Prove the distributive properties in Theorem 1.4.10.

3. Prove De Morgan's laws (Theorem 1.4.9).

4. Prove Theorem 1.4.11.

5. For every collection of events A_i $(i \in I)$, show that

$$\left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} A_i^c \quad \text{and} \quad \left(\bigcap_{i\in I} A_i\right)^c = \bigcup_{i\in I} A_i^c.$$

6. Suppose that one card is to be selected from a deck of 20 cards that contains 10 red cards numbered from 1 to 10 and 10 blue cards numbered from 1 to 10. Let A be the event that a card with an even number is selected, let B be the event that a blue card is selected, and let C be the event that a card with a number less than 5 is selected. Describe the sample space S and describe each of the following events both in words and as subsets of S:

a. $A \cap B \cap C$ **b.** $B \cap C^c$ **c.** $A \cup B \cup C$
d. $A \cap (B \cup C)$ **e.** $A^c \cap B^c \cap C^c$.

7. Suppose that a number x is to be selected from the real line S, and let A, B, and C be the events represented by the following subsets of S, where the notation $\{x\colon \text{- - -}\}$ denotes the set containing every point x for which the property presented following the colon is satisfied:

$$A = \{x\colon 1 \le x \le 5\},$$
$$B = \{x\colon 3 < x \le 7\},$$
$$C = \{x\colon x \le 0\}.$$

Describe each of the following events as a set of real numbers:

a. A^c **b.** $A \cup B$ **c.** $B \cap C^c$
d. $A^c \cap B^c \cap C^c$ **e.** $(A \cup B) \cap C$.

8. A simplified model of the human blood-type system has four blood types: A, B, AB, and O. There are two antigens, anti-A and anti-B, that react with a person's blood in different ways depending on the blood type. Anti-A reacts with blood types A and AB, but not with B and O. Anti-B reacts with blood types B and AB, but not with A and O. Suppose that a person's blood is sampled and tested with the two antigens. Let A be the event that the blood reacts with anti-A, and let B be the event that it reacts with anti-B. Classify the person's blood type using the events A, B, and their complements.

9. Let S be a given sample space and let $A_1, A_2, \ldots$ be an infinite sequence of events. For $n = 1, 2, \ldots$, let $B_n = \bigcup_{i=n}^{\infty} A_i$ and let $C_n = \bigcap_{i=n}^{\infty} A_i$.

a. Show that $B_1 \supset B_2 \supset \cdots$ and that $C_1 \subset C_2 \subset \cdots$.

b. Show that an outcome in S belongs to the event $\bigcap_{n=1}^{\infty} B_n$ if and only if it belongs to an infinite number of the events $A_1, A_2, \ldots$.

c. Show that an outcome in S belongs to the event $\bigcup_{n=1}^{\infty} C_n$ if and only if it belongs to all the events $A_1, A_2, \ldots$ except possibly a finite number of those events.

10. Three six-sided dice are rolled. The six sides of each die are numbered 1–6. Let A be the event that the first die shows an even number, let B be the event that the second die shows an even number, and let C be the event that the third die shows an even number. Also, for each $i = 1, \ldots, 6$, let A_i be the event that the first die shows the number i, let B_i be the event that the second die shows the number i, and let C_i be the event that the third die shows the number i. Express each of the following events in terms of the named events described above:

a. The event that all three dice show even numbers

b. The event that no die shows an even number

c. The event that at least one die shows an odd number

d. The event that at most two dice show odd numbers

e. The event that the sum of the three dices is no greater than 5

11. A power cell consists of two subcells, each of which can provide from 0 to 5 volts, regardless of what the other

subcell provides. The power cell is functional if and only if the sum of the two voltages of the subcells is at least 6 volts. An experiment consists of measuring and recording the voltages of the two subcells. Let A be the event that the power cell is functional, let B be the event that two subcells have the same voltage, let C be the event that the first subcell has a strictly higher voltage than the second subcell, and let D be the event that the power cell is not functional but needs less than one additional volt to become functional.

a. Define a sample space S for the experiment as a set of ordered pairs that makes it possible for you to express the four sets above as events.

b. Express each of the events A, B, C, and D as sets of ordered pairs that are subsets of S.

c. Express the following set in terms of A, B, C, and/or D: $\{(x, y) : x = y \text{ and } x + y \le 5\}$.

d. Express the following event in terms of A, B, C, and/or D: the event that the power cell is not functional and the second subcell has a strictly higher voltage than the first subcell.

12. Suppose that the sample space S of some experiment is finite. Show that the collection of all subsets of S satisfies the three conditions required to be called the collection of events.

13. Let S be the sample space for some experiment. Show that the collection of subsets consisting solely of S and $\emptyset$ satisfies the three conditions required in order to be called the collection of events. Explain why this collection would not be very interesting in most real problems.

14. Suppose that the sample space S of some experiment is countable. Suppose also that, for every outcome $s \in S$, the subset $\{s\}$ is an event. Show that every subset of S must be an event. *Hint:* Recall the three conditions required of the collection of subsets of S that we call events.

1.5 The Definition of Probability

We begin with the mathematical definition of probability and then present some useful results that follow easily from the definition.

Axioms and Basic Theorems

In this section, we shall present the mathematical, or axiomatic, definition of probability. In a given experiment, it is necessary to assign to each event A in the sample space S a number $\Pr(A)$ that indicates the probability that A will occur. In order to satisfy the mathematical definition of probability, the number $\Pr(A)$ that is assigned must satisfy three specific axioms. These axioms ensure that the number $\Pr(A)$ will have certain properties that we intuitively expect a probability to have under each of the various interpretations described in Sec. 1.2.

The first axiom states that the probability of every event must be nonnegative.

Axiom 1 For every event A, $\Pr(A) \ge 0$.

The second axiom states that if an event is certain to occur, then the probability of that event is 1.

Axiom 2 $\Pr(S) = 1$.

Before stating Axiom 3, we shall discuss the probabilities of disjoint events. If two events are disjoint, it is natural to assume that the probability that one or the other will occur is the sum of their individual probabilities. In fact, it will be assumed that this *additive property* of probability is also true for every finite collection of disjoint events and even for every infinite sequence of disjoint events. If we assume that this additive property is true only for a finite number of disjoint events, we cannot then be certain that the property will be true for an infinite sequence of disjoint events as well. However, if we assume that the additive property is true for every infinite sequence

of disjoint events, then (as we shall prove) the property must also be true for every finite number of disjoint events. These considerations lead to the third axiom.

Axiom 3 For every infinite sequence of disjoint events $A_1, A_2, \ldots,$

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i).$$

Example 1.5.1 Rolling a Die. In Example 1.4.1, for each subset A of $S = \{1, 2, 3, 4, 5, 6\}$, let $\Pr(A)$ be the number of elements of A divided by 6. It is trivial to see that this satisfies the first two axioms. There are only finitely many distinct collections of nonempty disjoint events. It is not difficult to see that Axiom 3 is also satisfied by this example. ◂

Example 1.5.2 A Loaded Die. In Example 1.5.1, there are other choices for the probabilities of events. For example, if we believe that the die is loaded, we might believe that some sides have different probabilities of turning up. To be specific, suppose that we believe that 6 is twice as likely to come up as each of the other five sides. We could set $p_i = 1/7$ for $i = 1, 2, 3, 4, 5$ and $p_6 = 2/7$. Then, for each event A, define $\Pr(A)$ to be the sum of all p_i such that $i \in A$. For example, if $A = \{1, 3, 5\}$, then $\Pr(A) = p_1 + p_3 + p_5 = 3/7$. It is not difficult to check that this also satisfies all three axioms. ◂

We are now prepared to give the mathematical definition of probability.

Definition 1.5.1 Probability. A *probability measure*, or simply a *probability*, on a sample space S is a specification of numbers $\Pr(A)$ for all events A that satisfy Axioms 1, 2, and 3.

We shall now derive two important consequences of Axiom 3. First, we shall show that if an event is impossible, its probability must be 0.

Theorem 1.5.1 $\Pr(\emptyset) = 0$.

Proof Consider the infinite sequence of events $A_1, A_2, \ldots$ such that $A_i = \emptyset$ for $i = 1, 2, \ldots$. In other words, each of the events in the sequence is just the empty set $\emptyset$. Then this sequence is a sequence of disjoint events, since $\emptyset \cap \emptyset = \emptyset$. Furthermore, $\bigcup_{i=1}^{\infty} A_i = \emptyset$. Therefore, it follows from Axiom 3 that

$$\Pr(\emptyset) = \Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i) = \sum_{i=1}^{\infty} \Pr(\emptyset).$$

This equation states that when the number $\Pr(\emptyset)$ is added repeatedly in an infinite series, the sum of that series is simply the number $\Pr(\emptyset)$. The only real number with this property is zero. ■

We can now show that the additive property assumed in Axiom 3 for an infinite sequence of disjoint events is also true for every finite number of disjoint events.

Theorem 1.5.2 For every finite sequence of n disjoint events $A_1, \ldots, A_n$,

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \Pr(A_i).$$

Proof Consider the infinite sequence of events $A_1, A_2, \ldots,$ in which $A_1, \ldots, A_n$ are the n given disjoint events and $A_i = \emptyset$ for $i > n$. Then the events in this infinite

sequence are disjoint and $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{n} A_i$. Therefore, by Axiom 3,

$$\begin{aligned}\Pr\left(\bigcup_{i=1}^{n} A_i\right) &= \Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)\\ &= \sum_{i=1}^{n} \Pr(A_i) + \sum_{i=n+1}^{\infty} \Pr(A_i)\\ &= \sum_{i=1}^{n} \Pr(A_i) + 0\\ &= \sum_{i=1}^{n} \Pr(A_i).\end{aligned}$$ ■

Further Properties of Probability

From the axioms and theorems just given, we shall now derive four other general properties of probability measures. Because of the fundamental nature of these four properties, they will be presented in the form of four theorems, each one of which is easily proved.

Theorem 1.5.3 For every event A, $\Pr(A^c) = 1 - \Pr(A)$.

Proof Since A and A^c are disjoint events and $A \cup A^c = S$, it follows from Theorem 1.5.2 that $\Pr(S) = \Pr(A) + \Pr(A^c)$. Since $\Pr(S) = 1$ by Axiom 2, then $\Pr(A^c) = 1 - \Pr(A)$. ■

Theorem 1.5.4 If $A \subset B$, then $\Pr(A) \leq \Pr(B)$.

Proof As illustrated in Fig. 1.8, the event B may be treated as the union of the two disjoint events A and $B \cap A^c$. Therefore, $\Pr(B) = \Pr(A) + \Pr(B \cap A^c)$. Since $\Pr(B \cap A^c) \geq 0$, then $\Pr(B) \geq \Pr(A)$. ■

Theorem 1.5.5 For every event A, $0 \leq \Pr(A) \leq 1$.

Proof It is known from Axiom 1 that $\Pr(A) \geq 0$. Since $A \subset S$ for every event A, Theorem 1.5.4 implies $\Pr(A) \leq \Pr(S) = 1$, by Axiom 2. ■

Theorem 1.5.6 For every two events A and B,

$$\Pr(A \cap B^c) = \Pr(A) - \Pr(A \cap B).$$

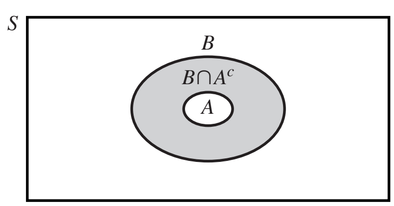


Figure 1.8 $B = A \cup (B \cap A^c)$ in the proof of Theorem 1.5.4.

Proof According to Theorem 1.4.11, the events $A \cap B^c$ and $A \cap B$ are disjoint and

$$A = (A \cap B) \cup (A \cap B^c).$$

It follows from Theorem 1.5.2 that

$$\Pr(A) = \Pr(A \cap B) + \Pr(A \cap B^c).$$

Subtract $\Pr(A \cap B)$ from both sides of this last equation to complete the proof. ■

Theorem 1.5.7 For every two events A and B,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \tag{1.5.1}$$

Proof From Theorem 1.4.11, we have

$$A \cup B = B \cup (A \cap B^c),$$

and the two events on the right side of this equation are disjoint. Hence, we have

$$\begin{aligned}\Pr(A \cup B) &= \Pr(B) + \Pr(A \cap B^c)\\ &= \Pr(B) + \Pr(A) - \Pr(A \cap B),\end{aligned}$$

where the first equation follows from Theorem 1.5.2, and the second follows from Theorem 1.5.6. ■

Example 1.5.3 Diagnosing Diseases. A patient arrives at a doctor's office with a sore throat and low-grade fever. After an exam, the doctor decides that the patient has either a bacterial infection or a viral infection or both. The doctor decides that there is a probability of 0.7 that the patient has a bacterial infection and a probability of 0.4 that the person has a viral infection. What is the probability that the patient has both infections?

Let B be the event that the patient has a bacterial infection, and let V be the event that the patient has a viral infection. We are told $\Pr(B) = 0.7$, that $\Pr(V) = 0.4$, and that $S = B \cup V$. We are asked to find $\Pr(B \cap V)$. We will use Theorem 1.5.7, which says that

$$\Pr(B \cup V) = \Pr(B) + \Pr(V) - \Pr(B \cap V). \tag{1.5.2}$$

Since $S = B \cup V$, the left-hand side of (1.5.2) is 1, while the first two terms on the right-hand side are 0.7 and 0.4. The result is

$$1 = 0.7 + 0.4 - \Pr(B \cap V),$$

which leads to $\Pr(B \cap V) = 0.1$, the probability that the patient has both infections. ◀

Example 1.5.4 Demands for Utilities. Consider, once again, the contractor who needs to plan for water and electricity demands in Example 1.4.5. There are many possible choices for how to spread the probability around the sample space (pictured in Fig. 1.5 on page 12). One simple choice is to make the probability of an event E proportional to the area of E. The area of S (the sample space) is $(150 - 1) \times (200 - 4) = 29{,}204$, so $\Pr(E)$ equals the area of E divided by 29,204. For example, suppose that the contractor is interested in high demand. Let A be the set where water demand is at least 100, and let B be the event that electric demand is at least 115, and suppose that these values are considered high demand. These events are shaded with different patterns in Fig. 1.9. The area of A is $(150 - 1) \times (200 - 100) = 14{,}900$, and the area

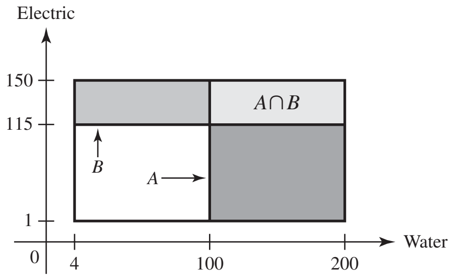


Figure 1.9 The two events of interest in utility demand sample space for Example 1.5.4.

of B is $(150 - 115) \times (200 - 4) = 6{,}860$. So,

$$\Pr(A) = \frac{14{,}900}{29{,}204} = 0.5102, \quad \Pr(B) = \frac{6{,}860}{29{,}204} = 0.2349.$$

The two events intersect in the region denoted by $A \cap B$. The area of this region is $(150 - 115) \times (200 - 100) = 3{,}500$, so $\Pr(A \cap B) = 3{,}500/29{,}204 = 0.1198$. If the contractor wishes to compute the probability that at least one of the two demands will be high, that probability is

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = 0.5102 + 0.2349 - 0.1198 = 0.6253,$$

according to Theorem 1.5.7. ◄

The proof of the following useful result is left to Exercise 13.

Theorem 1.5.8 Bonferroni Inequality. For all events $A_1, \ldots, A_n$,

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n} \Pr(A_i) \text{ and } \Pr\left(\bigcap_{i=1}^{n} A_i\right) \ge 1 - \sum_{i=1}^{n} \Pr(A_i^c).$$

(The second inequality above is known as the *Bonferroni inequality*.) ■

Note: Probability Zero Does Not Mean Impossible. When an event has probability 0, it does not mean that the event is impossible. In Example 1.5.4, there are many events with 0 probability, but they are not all impossible. For example, for every x, the event that water demand equals x corresponds to a line segment in Fig. 1.5. Since line segments have 0 area, the probability of every such line segment is 0, but the events are not all impossible. Indeed, if every event of the form {water demand equals x} were impossible, then water demand could not take any value at all. If $\epsilon > 0$, the event

$$\{\text{water demand is between } x - \epsilon \text{ and } x + \epsilon\}$$

will have positive probability, but that probability will go to 0 as ϵ goes to 0.

Summary

We have presented the mathematical definition of probability through the three axioms. The axioms require that every event have nonnegative probability, that the whole sample space have probability 1, and that the union of an infinite sequence of disjoint events have probability equal to the sum of their probabilities. Some important results to remember include the following:

- If $A_1, \dots, A_k$ are disjoint, $\Pr\left(\cup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \Pr(A_i)$.
- $\Pr(A^c) = 1 - \Pr(A)$.
- $A \subset B$ implies that $\Pr(A) \le \Pr(B)$.
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

It does not matter how the probabilities were determined. As long as they satisfy the three axioms, they must also satisfy the above relations as well as all of the results that we prove later in the text.

Exercises

1. One ball is to be selected from a box containing red, white, blue, yellow, and green balls. If the probability that the selected ball will be red is 1/5 and the probability that it will be white is 2/5, what is the probability that it will be blue, yellow, or green?

2. A student selected from a class will be either a boy or a girl. If the probability that a boy will be selected is 0.3, what is the probability that a girl will be selected?

3. Consider two events A and B such that $\Pr(A) = 1/3$ and $\Pr(B) = 1/2$. Determine the value of $\Pr(B \cap A^c)$ for each of the following conditions: **(a)** A and B are disjoint; **(b)** $A \subset B$; **(c)** $\Pr(A \cap B) = 1/8$.

4. If the probability that student A will fail a certain statistics examination is 0.5, the probability that student B will fail the examination is 0.2, and the probability that both student A and student B will fail the examination is 0.1, what is the probability that at least one of these two students will fail the examination?

5. For the conditions of Exercise 4, what is the probability that neither student A nor student B will fail the examination?

6. For the conditions of Exercise 4, what is the probability that exactly one of the two students will fail the examination?

7. Consider two events A and B with $\Pr(A) = 0.4$ and $\Pr(B) = 0.7$. Determine the maximum and minimum possible values of $\Pr(A \cap B)$ and the conditions under which each of these values is attained.

8. If 50 percent of the families in a certain city subscribe to the morning newspaper, 65 percent of the families subscribe to the afternoon newspaper, and 85 percent of the families subscribe to at least one of the two newspapers, what percentage of the families subscribe to both newspapers?

9. Prove that for every two events A and B, the probability that exactly one of the two events will occur is given by the expression

$$\Pr(A) + \Pr(B) - 2\Pr(A \cap B).$$

10. For two arbitrary events A and B, prove that

$$\Pr(A) = \Pr(A \cap B) + \Pr(A \cap B^c).$$

11. A point (x, y) is to be selected from the square S containing all points (x, y) such that $0 \le x \le 1$ and $0 \le y \le 1$. Suppose that the probability that the selected point will belong to each specified subset of S is equal to the area of that subset. Find the probability of each of the following subsets: **(a)** the subset of points such that $(x - \frac{1}{2})^2 + (y - \frac{1}{2})^2 \ge \frac{1}{4}$; **(b)** the subset of points such that $\frac{1}{2} < x + y < \frac{3}{2}$; **(c)** the subset of points such that $y \le 1 - x^2$; **(d)** the subset of points such that $x = y$.

12. Let $A_1, A_2, \dots$ be an arbitrary infinite sequence of events, and let $B_1, B_2, \dots$ be another infinite sequence of events defined as follows: $B_1 = A_1$, $B_2 = A_1^c \cap A_2$, $B_3 = A_1^c \cap A_2^c \cap A_3$, $B_4 = A_1^c \cap A_2^c \cap A_3^c \cap A_4, \dots$. Prove that

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \Pr(B_i) \text{ for } n = 1, 2, \dots,$$

and that

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(B_i).$$

13. Prove Theorem 1.5.8. *Hint:* Use Exercise 12.

14. Consider, once again, the four blood types A, B, AB, and O described in Exercise 8 in Sec. 1.4 together with the two antigens anti-A and anti-B. Suppose that, for a given person, the probability of type O blood is 0.5, the probability of type A blood is 0.34, and the probability of type B blood is 0.12.

a. Find the probability that each of the antigens will react with this person's blood.

b. Find the probability that both antigens will react with this person's blood.

1.6 Finite Sample Spaces

The simplest experiments in which to determine and derive probabilities are those that involve only finitely many possible outcomes. This section gives several examples to illustrate the important concepts from Sec. 1.5 in finite sample spaces.

Example 1.6.1 Current Population Survey. Every month, the Census Bureau conducts a survey of the United States population in order to learn about labor-force characteristics. Several pieces of information are collected on each of about 50,000 households. One piece of information is whether or not someone in the household is actively looking for employment but currently not employed. Suppose that our experiment consists of selecting three households at random from the 50,000 that were surveyed in a particular month and obtaining access to the information recorded during the survey. (Due to the confidential nature of information obtained during the Current Population Survey, only researchers in the Census Bureau would be able to perform the experiment just described.) The outcomes that make up the sample space S for this experiment can be described as lists of three three distinct numbers from 1 to 50,000. For example (300, 1, 24602) is one such list where we have kept track of the order in which the three households were selected. Clearly, there are only finitely many such lists. We can assume that each list is equally likely to be chosen, but we need to be able to count how many such lists there are. We shall learn a method for counting the outcomes for this example in Sec. 1.7. ◀

Requirements of Probabilities

In this section, we shall consider experiments for which there are only a finite number of possible outcomes. In other words, we shall consider experiments for which the sample space S contains only a finite number of points $s_1, \ldots, s_n$. In an experiment of this type, a probability measure on S is specified by assigning a probability p_i to each point $s_i \in S$. The number p_i is the probability that the outcome of the experiment will be s_i $(i = 1, \ldots, n)$. In order to satisfy the axioms of probability, the numbers $p_1, \ldots, p_n$ must satisfy the following two conditions:

$$p_i \geq 0 \quad \text{for } i = 1, \ldots, n$$

and

$$\sum_{i=1}^{n} p_i = 1.$$

The probability of each event A can then be found by adding the probabilities p_i of all outcomes s_i that belong to A. This is the general version of Example 1.5.2.

Example 1.6.2 Fiber Breaks. Consider an experiment in which five fibers having different lengths are subjected to a testing process to learn which fiber will break first. Suppose that the lengths of the five fibers are 1, 2, 3, 4, and 5 inches, respectively. Suppose also that the probability that any given fiber will be the first to break is proportional to the length of that fiber. We shall determine the probability that the length of the fiber that breaks first is not more than 3 inches.

In this example, we shall let s_i be the outcome in which the fiber whose length is i inches breaks first $(i = 1, \ldots, 5)$. Then $S = \{s_1, \ldots, s_5\}$ and $p_i = \alpha i$ for $i = 1, \ldots, 5$, where α is a proportionality factor. It must be true that $p_1 + \cdots + p_5 = 1$, and we know that $p_1 + \cdots + p_5 = 15\alpha$, so $\alpha = 1/15$. If A is the event that the length of the

fiber that breaks first is not more than 3 inches, then $A = \{s_1, s_2, s_3\}$. Therefore,

$$\Pr(A) = p_1 + p_2 + p_3 = \frac{1}{15} + \frac{2}{15} + \frac{3}{15} = \frac{2}{5}. \quad \blacktriangleleft$$

Simple Sample Spaces

A sample space S containing n outcomes $s_1, \ldots, s_n$ is called a simple sample space if the probability assigned to each of the outcomes $s_1, \ldots, s_n$ is $1/n$. If an event A in this simple sample space contains exactly m outcomes, then

$$\Pr(A) = \frac{m}{n}.$$

Example 1.6.3 Tossing Coins. Suppose that three fair coins are tossed simultaneously. We shall determine the probability of obtaining exactly two heads.

Regardless of whether or not the three coins can be distinguished from each other by the experimenter, it is convenient for the purpose of describing the sample space to assume that the coins can be distinguished. We can then speak of the result for the first coin, the result for the second coin, and the result for the third coin; and the sample space will comprise the eight possible outcomes listed in Example 1.4.4 on page 12.

Furthermore, because of the assumption that the coins are fair, it is reasonable to assume that this sample space is simple and that the probability assigned to each of the eight outcomes is 1/8. As can be seen from the listing in Example 1.4.4, exactly two heads will be obtained in three of these outcomes. Therefore, the probability of obtaining exactly two heads is 3/8. ◀

It should be noted that if we had considered the only possible outcomes to be no heads, one head, two heads, and three heads, it would have been reasonable to assume that the sample space contained just these four outcomes. This sample space would not be simple because the outcomes *would not be equally probable*.

Example 1.6.4 Genetics. Inherited traits in humans are determined by material in specific locations on chromosomes. Each normal human receives 23 chromosomes from each parent, and these chromosomes are naturally paired, with one chromosome in each pair coming from each parent. For the purposes of this text, it is safe to think of a *gene* as a portion of each chromosome in a pair. The genes, either one at a time or in combination, determine the inherited traits, such as blood type and hair color. The material in the two locations that make up a gene on the pair of chromosomes comes in forms called *alleles*. Each distinct combination of alleles (one on each chromosome) is called a *genotype*.

Consider a gene with only two different alleles A and a. Suppose that both parents have genotype Aa, that is, each parent has allele A on one chromosome and allele a on the other. (We do not distinguish the same alleles in a different order as a different genotype. For example, aA would be the same genotype as Aa. But it can be convenient to distinguish the two chromosomes during intermediate steps in probability calculations, just as we distinguished the three coins in Example 1.6.3.) What are the possible genotypes of an offspring of these two parents? If all possible results of the parents contributing pairs of alleles are equally likely, what are the probabilities of the different genotypes?

To begin, we shall distinguish which allele the offspring receives from each parent, since we are assuming that pairs of contributed alleles are equally likely.

Afterward, we shall combine those results that produce the same genotype. The possible contributions from the parents are:

	Mother	
Father	A	a
A	AA	Aa
a	aA	aa

So, there are three possible genotypes AA, Aa, and aa for the offspring. Since we assumed that every combination was equally likely, the four cells in the table all have probability 1/4. Since two of the cells in the table combined into genotype Aa, that genotype has probability 1/2. The other two genotypes each have probability 1/4, since they each correspond to only one cell in the table. ◀

Example 1.6.5 Rolling Two Dice. We shall now consider an experiment in which two balanced dice are rolled, and we shall calculate the probability of each of the possible values of the sum of the two numbers that may appear.

Although the experimenter need not be able to distinguish the two dice from one another in order to observe the value of their sum, the specification of a simple sample space in this example will be facilitated if we assume that the two dice are distinguishable. If this assumption is made, each outcome in the sample space S can be represented as a pair of numbers (x, y), where x is the number that appears on the first die and y is the number that appears on the second die. Therefore, S comprises the following 36 outcomes:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

It is natural to assume that S is a simple sample space and that the probability of each of these outcomes is 1/36.

Let P_i denote the probability that the sum of the two numbers is i for $i = 2, 3, \ldots, 12$. The only outcome in S for which the sum is 2 is the outcome (1, 1). Therefore, $P_2 = 1/36$. The sum will be 3 for either of the two outcomes (1, 2) and (2, 1). Therefore, $P_3 = 2/36 = 1/18$. By continuing in this manner, we obtain the following probability for each of the possible values of the sum:

$$
\begin{aligned}
P_2 &= P_{12} = \frac{1}{36}, & P_5 &= P_9 = \frac{4}{36}, \\
P_3 &= P_{11} = \frac{2}{36}, & P_6 &= P_8 = \frac{5}{36}, \\
P_4 &= P_{10} = \frac{3}{36}, & P_7 &= \frac{6}{36}.
\end{aligned}
$$

◀

Summary

A simple sample space is a finite sample space S such that every outcome in S has the same probability. If there are n outcomes in a simple sample space S, then each one must have probability $1/n$. The probability of an event E in a simple sample space is the number of outcomes in E divided by n. In the next three sections, we will present some useful methods for counting numbers of outcomes in various events.

Exercises

1. If two balanced dice are rolled, what is the probability that the sum of the two numbers that appear will be odd?

2. If two balanced dice are rolled, what is the probability that the sum of the two numbers that appear will be even?

3. If two balanced dice are rolled, what is the probability that the difference between the two numbers that appear will be less than 3?

4. A school contains students in grades 1, 2, 3, 4, 5, and 6. Grades 2, 3, 4, 5, and 6 all contain the same number of students, but there are twice this number in grade 1. If a student is selected at random from a list of all the students in the school, what is the probability that she will be in grade 3?

5. For the conditions of Exercise 4, what is the probability that the selected student will be in an odd-numbered grade?

6. If three fair coins are tossed, what is the probability that all three faces will be the same?

7. Consider the setup of Example 1.6.4 on page 23. This time, assume that two parents have genotypes Aa and aa. Find the possible genotypes for an offspring and find the probabilities for each genotype. Assume that all possible results of the parents contributing pairs of alleles are equally likely.

8. Consider an experiment in which a fair coin is tossed once and a balanced die is rolled once.

a. Describe the sample space for this experiment.

b. What is the probability that a head will be obtained on the coin and an odd number will be obtained on the die?

1.7 Counting Methods

In simple sample spaces, one way to calculate the probability of an event involves counting the number of outcomes in the event and the number of outcomes in the sample space. This section presents some common methods for counting the number of outcomes in a set. These methods rely on special structure that exists in many common experiments, namely, that each outcome consists of several parts and that it is relatively easy to count how many possibilities there are for each of the parts.

We have seen that in a simple sample space S, the probability of an event A is the ratio of the number of outcomes in A to the total number of outcomes in S. In many experiments, the number of outcomes in S is so large that a complete listing of these outcomes is too expensive, too slow, or too likely to be incorrect to be useful. In such an experiment, it is convenient to have a method of determining the total number of outcomes in the space S and in various events in S without compiling a list of all these outcomes. In this section, some of these methods will be presented.

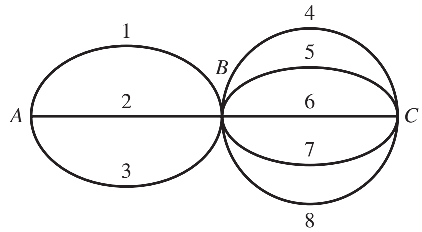


Figure 1.10 Three cities with routes between them in Example 1.7.1.

Multiplication Rule

Example 1.7.1 Routes between Cities. Suppose that there are three different routes from city A to city B and five different routes from city B to city C. The cities and routes are depicted in Fig. 1.10, with the routes numbered from 1 to 8. We wish to count the number of different routes from A to C that pass through B. For example, one such route from Fig. 1.10 is 1 followed by 4, which we can denote (1, 4). Similarly, there are the routes (1, 5), (1, 6), . . . , (3, 8). It is not difficult to see that the number of different routes $3 \times 5 = 15$. ◀

Example 1.7.1 is a special case of a common form of experiment.

Example 1.7.2 Experiment in Two Parts. Consider an experiment that has the following two characteristics:

i. The experiment is performed in two parts.
ii. The first part of the experiment has m possible outcomes $x_1, \dots, x_m$, and, regardless of which one of these outcomes x_i occurs, the second part of the experiment has n possible outcomes $y_1, \dots, y_n$.

Each outcome in the sample space S of such an experiment will therefore be a pair having the form (x_i, y_j), and S will be composed of the following pairs:

$$\begin{array}{l} (x_1, y_1)(x_1, y_2) \cdots (x_1, y_n) \\ (x_2, y_1)(x_2, y_2) \cdots (x_2, y_n) \\ \quad\vdots \qquad\quad \vdots \qquad\qquad \vdots \\ (x_m, y_1)(x_m, y_2) \cdots (x_m, y_n). \end{array}$$ ◀

Since each of the m rows in the array in Example 1.7.2 contains n pairs, the following result follows directly.

Theorem 1.7.1 Multiplication Rule for Two-Part Experiments. In an experiment of the type described in Example 1.7.2, the sample space S contains exactly mn outcomes. ■

Figure 1.11 illustrates the multiplication rule for the case of $n = 3$ and $m = 2$ with a tree diagram. Each end-node of the tree represents an outcome, which is the pair consisting of the two parts whose names appear along the branch leading to the end-node.

Example 1.7.3 Rolling Two Dice. Suppose that two dice are rolled. Since there are six possible outcomes for each die, the number of possible outcomes for the experiment is $6 \times 6 = 36$, as we saw in Example 1.6.5. ◀

The multiplication rule can be extended to experiments with more than two parts.

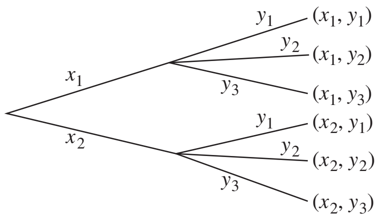


Figure 1.11 Tree diagram in which end-nodes represent outcomes.

Theorem 1.7.2 Multiplication Rule. Suppose that an experiment has k parts ($k \geq 2$), that the ith part of the experiment can have n_i possible outcomes ($i = 1, \dots, k$), and that all of the outcomes in each part can occur regardless of which specific outcomes have occurred in the other parts. Then the sample space S of the experiment will contain all vectors of the form $(u_1, \dots, u_k)$, where u_i is one of the n_i possible outcomes of part i ($i = 1, \dots, k$). The total number of these vectors in S will be equal to the product $n_1 n_2 \cdots n_k$. ■

Example 1.7.4 Tossing Several Coins. Suppose that we toss six coins. Each outcome in S will consist of a sequence of six heads and tails, such as HTTHHH. Since there are two possible outcomes for each of the six coins, the total number of outcomes in S will be $2^6 = 64$. If head and tail are considered equally likely for each coin, then S will be a simple sample space. Since there is only one outcome in S with six heads and no tails, the probability of obtaining heads on all six coins is 1/64. Since there are six outcomes in S with one head and five tails, the probability of obtaining exactly one head is $6/64 = 3/32$. ◀

Example 1.7.5 Combination Lock. A standard combination lock has a dial with tick marks for 40 numbers from 0 to 39. The combination consists of a sequence of three numbers that must be dialed in the correct order to open the lock. Each of the 40 numbers may appear in each of the three positions of the combination regardless of what the other two positions contain. It follows that there are $40^3 = 64{,}000$ possible combinations. This number is supposed to be large enough to discourage would-be thieves from trying every combination. ◀

Note: The Multiplication Rule Is Slightly More General. In the statements of Theorems 1.7.1 and 1.7.2, it is assumed that each possible outcome in each part of the experiment can occur regardless of what occurs in the other parts of the experiment. Technically, all that is necessary is that the *number* of possible outcomes for each part of the experiment not depend on what occurs on the other parts. The discussion of permutations below is an example of this situation.

Permutations

Example 1.7.6 Sampling without Replacement. Consider an experiment in which a card is selected and removed from a deck of n different cards, a second card is then selected and removed from the remaining $n - 1$ cards, and finally a third card is selected from the remaining $n - 2$ cards. Each outcome consists of the three cards in the order selected. A process of this kind is called *sampling without replacement*, since a card that is drawn is not replaced in the deck before the next card is selected. In this experiment, any one of the n cards could be selected first. Once this card has been removed, any one of the other $n - 1$ cards could be selected second. Therefore, there are $n(n - 1)$

possible outcomes for the first two selections. Finally, for every given outcome of the first two selections, there are $n - 2$ other cards that could possibly be selected third. Therefore, the total number of possible outcomes for all three selections is $n(n - 1)(n - 2)$. ◀

The situation in Example 1.7.6 can be generalized to any number of selections without replacement.

Definition 1.7.1 Permutations. Suppose that a set has n elements. Suppose that an experiment consists of selecting k of the elements one at a time without replacement. Let each outcome consist of the k elements in the order selected. Each such outcome is called a *permutation of n elements taken k at a time*. We denote the number of distinct such permutations by the symbol $P_{n,k}$.

By arguing as in Example 1.7.6, we can figure out how many different permutations there are of n elements taken k at a time. The proof of the following theorem is simply to extend the reasoning in Example 1.7.6 to selecting k cards without replacement. The proof is left to the reader.

Theorem 1.7.3 Number of Permutations. The number of permutations of n elements taken k at a time is $P_{n,k} = n(n - 1) \cdots (n - k + 1)$. ■

Example 1.7.7 Current Population Survey. Theorem 1.7.3 allows us to count the number of points in the sample space of Example 1.6.1. Each outcome in S consists of a permutation of $n = 50{,}000$ elements taken $k = 3$ at a time. Hence, the sample space S in that example consisits of

$$50{,}000 \times 49{,}999 \times 49{,}998 = 1.25 \times 10^{14}$$

outcomes. ◀

When $k = n$, the number of possible permutations will be the number $P_{n,n}$ of different permutations of all n cards. It is seen from the equation just derived that

$$P_{n,n} = n(n - 1) \cdots 1 = n!$$

The symbol $n!$ is read *n factorial*. In general, the number of permutations of n different items is $n!$.

The expression for $P_{n,k}$ can be rewritten in the following alternate form for $k = 1, \ldots, n - 1$:

$$P_{n,k} = n(n - 1) \cdots (n - k + 1)\frac{(n - k)(n - k - 1) \cdots 1}{(n - k)(n - k - 1) \cdots 1} = \frac{n!}{(n - k)!}.$$

Here and elsewhere in the theory of probability, it is convenient to define $0!$ by the relation

$$0! = 1.$$

With this definition, it follows that the relation $P_{n,k} = n!/(n - k)!$ will be correct for the value $k = n$ as well as for the values $k = 1, \ldots, n - 1$. To summarize:

Theorem 1.7.4 Permutations. The number of distinct orderings of k items selected without replacement from a collection of n different items $(0 \le k \le n)$ is

$$P_{n,k} = \frac{n!}{(n - k)!}.$$ ■

Example 1.7.8 Choosing Officers. Suppose that a club consists of 25 members and that a president and a secretary are to be chosen from the membership. We shall determine the total possible number of ways in which these two positions can be filled.

Since the positions can be filled by first choosing one of the 25 members to be president and then choosing one of the remaining 24 members to be secretary, the possible number of choices is $P_{25,2} = (25)(24) = 600$. ◀

Example 1.7.9 Arranging Books. Suppose that six different books are to be arranged on a shelf. The number of possible permutations of the books is $6! = 720$. ◀

Example 1.7.10 Sampling with Replacement. Consider a box that contains n balls numbered $1, \ldots, n$. First, one ball is selected at random from the box and its number is noted. This ball is then put back in the box and another ball is selected (it is possible that the same ball will be selected again). As many balls as desired can be selected in this way. This process is called *sampling with replacement*. It is assumed that each of the n balls is equally likely to be selected at each stage and that all selections are made independently of each other.

Suppose that a total of k selections are to be made, where k is a given positive integer. Then the sample space S of this experiment will contain all vectors of the form $(x_1, \ldots, x_k)$, where x_i is the outcome of the ith selection ($i = 1, \ldots, k$). Since there are n possible outcomes for each of the k selections, the total number of vectors in S is n^k. Furthermore, from our assumptions it follows that S is a simple sample space. Hence, the probability assigned to each vector in S is $1/n^k$. ◀

Example 1.7.11 Obtaining Different Numbers. For the experiment in Example 1.7.10, we shall determine the probability of the event E that each of the k balls that are selected will have a different number.

If $k > n$, it is impossible for all the selected balls to have different numbers because there are only n different numbers. Suppose, therefore, that $k \leq n$. The number of outcomes in the event E is the number of vectors for which all k components are different. This equals $P_{n,k}$, since the first component x_1 of each vector can have n possible values, the second component x_2 can then have any one of the other $n - 1$ values, and so on. Since S is a simple sample space containing n^k vectors, the probability p that k different numbers will be selected is

$$p = \frac{P_{n,k}}{n^k} = \frac{n!}{(n-k)!n^k}.$$ ◀

Note: Using Two Different Methods in the Same Problem. Example 1.7.11 illustrates a combination of techniques that might seem confusing at first. The method used to count the number of outcomes in the sample space was based on sampling with replacement, since the experiment allows repeat numbers in each outcome. The method used to count the number of outcomes in the event E was permutations (sampling without replacement) because E consists of those outcomes without repeats. It often happens that one needs to use different methods to count the numbers of outcomes in different subsets of the sample space. The birthday problem, which follows, is another example in which we need more than one counting method in the same problem.

The Birthday Problem

In the following problem, which is often called the birthday problem, it is required to determine the probability p that at least two people in a group of k people will have the same birthday, that is, will have been born on the same day of the same month but not necessarily in the same year. For the solution presented here, we assume that the birthdays of the k people are unrelated (in particular, we assume that twins are not present) and that each of the 365 days of the year is equally likely to be the birthday of any person in the group. In particular, we ignore the fact that the birth rate actually varies during the year and we assume that anyone actually born on February 29 will consider his birthday to be another day, such as March 1.

When these assumptions are made, this problem becomes similar to the one in Example 1.7.11. Since there are 365 possible birthdays for each of k people, the sample space S will contain 365^k outcomes, all of which will be equally probable. If $k > 365$, there are not enough birthdays for every one to be different, and hence at least two people *must* have the same birthday. So, we assume that $k \leq 365$. Counting the number of outcomes in which at least two birthdays are the same is tedious. However, the number of outcomes in S for which all k birthdays will be different is $P_{365,k}$, since the first person's birthday could be any one of the 365 days, the second person's birthday could then be any of the other 364 days, and so on. Hence, the probability that all k persons will have different birthdays is

$$\frac{P_{365,k}}{365^k}.$$

The probability p that at least two of the people will have the same birthday is therefore

$$p = 1 - \frac{P_{365,k}}{365^k} = 1 - \frac{(365)!}{(365-k)!365^k}.$$

Numerical values of this probability p for various values of k are given in Table 1.1. These probabilities may seem surprisingly large to anyone who has not thought about them before. Many persons would guess that in order to obtain a value of p greater than 1/2, the number of people in the group would have to be about 100. However, according to Table 1.1, there would have to be only 23 people in the group. As a matter of fact, for $k = 100$ the value of p is 0.9999997.

Table 1.1 The probability p that at least two people in a group of k people will have the same birthday

k	p	k	p
5	0.027	25	0.569
10	0.117	30	0.706
15	0.253	40	0.891
20	0.411	50	0.970
22	0.476	60	0.994
23	0.507		

The calculation in this example illustrates a common technique for solving probability problems. If one wishes to compute the probability of some event A, it might be more straightforward to calculate $\Pr(A^c)$ and then use the fact that $\Pr(A) = 1 - \Pr(A^c)$. This idea is particularly useful when the event A is of the form "at least n things happen" where n is small compared to how many things could happen.

Stirling's Formula

For large values of n, it is nearly impossible to compute $n!$. For $n \geq 70$, $n! > 10^{100}$ and cannot be represented on many scientific calculators. In most cases for which $n!$ is needed with a large value of n, one only needs the ratio of $n!$ to another large number a_n. A common example of this is $P_{n,k}$ with large n and not so large k, which equals $n!/(n-k)!$. In such cases, we can notice that

$$\frac{n!}{a_n} = e^{\log(n!) - \log(a_n)}.$$

Compared to computing $n!$, it takes a much larger n before $\log(n!)$ becomes difficult to represent. Furthermore, if we had a simple approximation s_n to $\log(n!)$ such that $\lim_{n\to\infty} |s_n - \log(n!)| = 0$, then the ratio of $n!/a_n$ to s_n/a_n would be close to 1 for large n. The following result, whose proof can be found in Feller (1968), provides such an approximation.

Theorem 1.7.5 Stirling's Formula. Let

$$s_n = \frac{1}{2}\log(2\pi) + \left(n + \frac{1}{2}\right)\log(n) - n.$$

Then $\lim_{n\to\infty} |s_n - \log(n!)| = 0$. Put another way,

$$\lim_{n\to\infty} \frac{(2\pi)^{1/2} n^{n+1/2} e^{-n}}{n!} = 1. \qquad \blacksquare$$

Example 1.7.12 Approximating the Number of Permutations. Suppose that we want to compute $P_{70,20} = 70!/50!$. The approximation from Stirling's formula is

$$\frac{70!}{50!} \approx \frac{(2\pi)^{1/2} 70^{70.5} e^{-70}}{(2\pi)^{1/2} 50^{50.5} e^{-50}} = 3.940 \times 10^{35}.$$

The exact calculation yields 3.938×10^{35}. The approximation and the exact calculation differ by less than 1/10 of 1 percent. ◀

Summary

Suppose that the following conditions are met:

- Each element of a set consists of k distinguishable parts $x_1, \ldots, x_k$.
- There are n_1 possibilities for the first part x_1.
- For each $i = 2, \ldots, k$ and each combination $(x_1, \ldots, x_{i-1})$ of the first $i - 1$ parts, there are n_i possibilities for the ith part x_i.

Under these conditions, there are $n_1 \cdots n_k$ elements of the set. The third condition requires only that the number of possibilities for x_i be n_i no matter what the earlier

parts are. For example, for $i = 2$, it does *not* require that the same n_2 possibilities be available for x_2 regardless of what x_1 is. It only requires that the *number* of possibilities for x_2 be n_2 no matter what x_1 is. In this way, the general rule includes the multiplication rule, the calculation of permutations, and sampling with replacement as special cases. For permutations of m items k at a time, we have $n_i = m - i + 1$ for $i = 1, \ldots, k$, and the n_i possibilities for part i are just the n_i items that have not yet appeared in the first $i - 1$ parts. For sampling with replacement from m items, we have $n_i = m$ for all i, and the m possibilities are the same for every part. In the next section, we shall consider how to count elements of sets in which the parts of each element are not distinguishable.

Exercises

1. Each year starts on one of the seven days (Sunday through Saturday). Each year is either a leap year (i.e., it includes February 29) or not. How many different calendars are possible for a year?

2. Three different classes contain 20, 18, and 25 students, respectively, and no student is a member of more than one class. If a team is to be composed of one student from each of these three classes, in how many different ways can the members of the team be chosen?

3. In how many different ways can the five letters a, b, c, d, and e be arranged?

4. If a man has six different sportshirts and four different pairs of slacks, how many different combinations can he wear?

5. If four dice are rolled, what is the probability that each of the four numbers that appear will be different?

6. If six dice are rolled, what is the probability that each of the six different numbers will appear exactly once?

7. If 12 balls are thrown at random into 20 boxes, what is the probability that no box will receive more than one ball?

8. An elevator in a building starts with five passengers and stops at seven floors. If every passenger is equally likely to get off at each floor and all the passengers leave independently of each other, what is the probability that no two passengers will get off at the same floor?

9. Suppose that three runners from team A and three runners from team B participate in a race. If all six runners have equal ability and there are no ties, what is the probability that the three runners from team A will finish first, second, and third, and the three runners from team B will finish fourth, fifth, and sixth?

10. A box contains 100 balls, of which r are red. Suppose that the balls are drawn from the box one at a time, at random, without replacement. Determine **(a)** the probability that the first ball drawn will be red; **(b)** the probability that the 50th ball drawn will be red; and **(c)** the probability that the last ball drawn will be red.

11. Let n and k be positive integers such that both n and $n - k$ are large. Use Stirling's formula to write as simple an approximation as you can for $P_{n,k}$.

1.8 Combinatorial Methods

Many problems of counting the number of outcomes in an event amount to counting how many subsets of a certain size are contained in a fixed set. This section gives examples of how to do such counting and where it can arise.

Combinations

Example 1.8.1 Choosing Subsets. Consider the set $\{a, b, c, d\}$ containing the four different letters. We want to count the number of distinct subsets of size two. In this case, we can list all of the subsets of size two:

$$\{a, b\}, \quad \{a, c\}, \quad \{a, d\}, \quad \{b, c\}, \quad \{b, d\}, \quad \text{and} \quad \{c, d\}.$$

We see that there are six distinct subsets of size two. This is different from counting permutaions because $\{a, b\}$ and $\{b, a\}$ are the same subset. ◀

For large sets, it would be tedious, if not impossible, to enumerate all of the subsets of a given size and count them as we did in Example 1.8.1. However, there is a connection between counting subsets and counting permutations that will allow us to derive the general formula for the number of subsets.

Suppose that there is a set of n distinct elements from which it is desired to choose a subset containing k elements $(1 \le k \le n)$. We shall determine the number of different subsets that can be chosen. In this problem, the arrangement of the elements in a subset is irrelevant and each subset is treated as a unit.

Definition 1.8.1 Combinations. Consider a set with n elements. Each subset of size k chosen from this set is called a *combination of n elements taken k at a time*. We denote the number of distinct such combinations by the symbol $C_{n,k}$.

No two combinations will consist of exactly the same elements because two subsets with the same elements are the same subset.

At the end of Example 1.8.1, we noted that two different permutations (a, b) and (b, a) both correspond to the same combination or subset $\{a, b\}$. We can think of permutations as being constructed in two steps. First, a combination of k elements is chosen out of n, and second, those k elements are arranged in a specific order. There are $C_{n,k}$ ways to choose the k elements out of n, and for each such choice there are $k!$ ways to arrange those k elements in different orders. Using the multiplication rule from Sec. 1.7, we see that the number of permutations of n elements taken k at a time is $P_{n,k} = C_{n,k}k!$; hence, we have the following.

Theorem 1.8.1 Combinations. The number of distinct subsets of size k that can be chosen from a set of size n is

$$C_{n,k} = \frac{P_{n,k}}{k!} = \frac{n!}{k!(n-k)!}. \qquad \blacksquare$$

In Example 1.8.1, we see that $C_{4,2} = 4!/[2!2!] = 6$.

Example 1.8.2 Selecting a Committee. Suppose that a committee composed of eight people is to be selected from a group of 20 people. The number of different groups of people that might be on the committee is

$$C_{20,8} = \frac{20!}{8!12!} = 125{,}970.$$ ◀

Example 1.8.3 Choosing Jobs. Suppose that, in Example 1.8.2, the eight people in the committee each get a different job to perform on the committee. The number of ways to choose eight people out of 20 and assign them to the eight different jobs is the number of permutations of 20 elements taken eight at a time, or

$$P_{20,8} = C_{20,8} \times 8! = 125{,}970 \times 8! = 5{,}078{,}110{,}400.$$ ◀

Examples 1.8.2 and 1.8.3 illustrate the difference and relationship between combinations and permutations. In Example 1.8.3, we count the same group of people in a different order as a different outcome, while in Example 1.8.2, we count the same group in different orders as the same outcome. The two numerical values differ by a factor of 8!, the number of ways to reorder each of the combinations in Example 1.8.2 to get a permutation in Example 1.8.3.

Binomial Coefficients

Definition 1.8.2

Binomial Coefficients. The number $C_{n,k}$ is also denoted by the symbol $\binom{n}{k}$. That is, for $k = 0, 1, \ldots, n$,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \quad (1.8.1)$$

When this notation is used, this number is called a *binomial coefficient*.

The name *binomial coefficient* derives from the appearance of the symbol in the binomial theorem, whose proof is left as Exercise 20 in this section.

Theorem 1.8.2

Binomial Theorem. For all numbers x and y and each positive integer n,

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}. \quad \blacksquare$$

There are a couple of useful relations between binomial coefficients.

Theorem 1.8.3

For all n,

$$\binom{n}{0} = \binom{n}{n} = 1.$$

For all n and all $k = 0, 1, \ldots, n$,

$$\binom{n}{k} = \binom{n}{n-k}.$$

Proof The first equation follows from the fact that $0! = 1$. The second equation follows from Eq. (1.8.1). The second equation can also be derived from the fact that selecting k elements to form a subset is equivalent to selecting the remaining $n-k$ elements to form the complement of the subset. ■

It is sometimes convenient to use the expression "n choose k" for the value of $C_{n,k}$. Thus, the same quantity is represented by the two different notations $C_{n,k}$ and $\binom{n}{k}$, and we may refer to this quantity in three different ways: as the number of combinations of n elements taken k at a time, as the binomial coefficient of n and k, or simply as "n choose k."

Example 1.8.4

Blood Types. In Example 1.6.4 on page 23, we defined genes, alleles, and genotypes. The gene for human blood type consists of a pair of alleles chosen from the three alleles commonly called O, A, and B. For example, two possible combinations of alleles (called genotypes) to form a blood-type gene would be BB and AO. We will not distinguish the same two alleles in different orders, so OA represents the same genotype as AO. How many genotypes are there for blood type?

The answer could easily be found by counting, but it is an example of a more general calculation. Suppose that a gene consists of a pair chosen from a set of n different alleles. Assuming that we cannot distinguish the same pair in different orders, there are n pairs where both alleles are the same, and there are $\binom{n}{2}$ pairs where the two alleles are different. The total number of genotypes is

$$n + \binom{n}{2} = n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2} = \binom{n+1}{2}.$$

For the case of blood type, we have $n = 3$, so there are

$$\binom{4}{2} = \frac{4 \times 3}{2} = 6$$

genotypes, as could easily be verified by counting. ◀

Note: Sampling with Replacement. The counting method described in Example 1.8.4 is a type of sampling with replacement that is different from the type described in Example 1.7.10. In Example 1.7.10, we sampled with replacement, but we distinguished between samples having the same balls in different orders. This could be called *ordered sampling with replacement*. In Example 1.8.4, samples containing the same genes in different orders were considered the same outcome. This could be called *unordered sampling with replacement*. The general formula for the number of unordered samples of size k with replacement from n elements is $\binom{n+k-1}{k}$, and can be derived in Exercise 19. It is possible to have k larger than n when sampling with replacement.

Example 1.8.5

Selecting Baked Goods. You go to a bakery to select some baked goods for a dinner party. You need to choose a total of 12 items. The baker has seven different types of items from which to choose, with lots of each type available. How many different boxfuls of 12 items are possible for you to choose? Here we will not distinguish the same collection of 12 items arranged in different orders in the box. This is an example of unordered sampling with replacement because we can (indeed we must) choose the same type of item more than once, but we are not distinguishing the same items in different orders. There are $\binom{7+12-1}{12} = 18{,}564$ different boxfuls. ◀

Example 1.8.5 raises an issue that can cause confusion if one does not carefully determine the elements of the sample space and carefully specify which outcomes (if any) are equally likely. The next example illustrates the issue in the context of Example 1.8.5.

Example 1.8.6

Selecting Baked Goods. Imagine two different ways of choosing a boxful of 12 baked goods selected from the seven different types available. In the first method, you choose one item at random from the seven available. Then, without regard to what item was chosen first, you choose the second item at random from the seven available. Then you continue in this way choosing the next item at random from the seven available without regard to what has already been chosen until you have chosen 12. For this method of choosing, it is natural to let the outcomes be the possible sequences of the 12 types of items chosen. The sample space would contain $7^{12} = 1.38 \times 10^{10}$ different outcomes that would be equally likely.

In the second method of choosing, the baker tells you that she has available 18,564 different boxfuls freshly packed. You then select one at random. In this case, the sample space would consist of 18,564 different equally likely outcomes.

In spite of the different sample spaces that arise in the two methods of choosing, there are some verbal descriptions that identify an event in both sample spaces. For example, both sample spaces contain an event that could be described as {all 12 items are of the same type} even though the outcomes are different types of mathematical objects in the two sample spaces. The probability that all 12 items are of the same type will actually be different depending on which method you use to choose the boxful.

In the first method, seven of the 7^{12} equally likely outcomes contain 12 of the same type of item. Hence, the probability that all 12 items are of the same type is

$7/7^{12} = 5.06 \times 10^{-10}$. In the second method, there are seven equally liklely boxes that contain 12 of the same type of item. Hence, the probability that all 12 items are of the same type is $7/18{,}564 = 3.77 \times 10^{-4}$. Before one can compute the probability for an event such as {all 12 items are of the same type}, one must be careful about defining the experiment and its outcomes. ◂

Arrangements of Elements of Two Distinct Types When a set contains only elements of two distinct types, a binomial coefficient can be used to represent the number of different arrangements of all the elements in the set. Suppose, for example, that k similar red balls and $n-k$ similar green balls are to be arranged in a row. Since the red balls will occupy k positions in the row, each different arrangement of the n balls corresponds to a different choice of the k positions occupied by the red balls. Hence, the number of different arrangements of the n balls will be equal to the number of different ways in which k positions can be selected for the red balls from the n available positions. Since this number of ways is specified by the binomial coefficient $\binom{n}{k}$, the number of different arrangements of the n balls is also $\binom{n}{k}$. In other words, the number of different arrangements of n objects consisting of k similar objects of one type and $n-k$ similar objects of a second type is $\binom{n}{k}$.

Example 1.8.7

Tossing a Coin. Suppose that a fair coin is to be tossed 10 times, and it is desired to determine (a) the probability p of obtaining exactly three heads and (b) the probability p' of obtaining three or fewer heads.

(a) The total possible number of different sequences of 10 heads and tails is 2^{10}, and it may be assumed that each of these sequences is equally probable. The number of these sequences that contain exactly three heads will be equal to the number of different arrangements that can be formed with three heads and seven tails. Here are some of those arrangements:

HHHTTTTTTT, HHTHTTTTTT, HHTTHTTTTT, TTHTHTHTTT, etc.

Each such arrangement is equivalent to a choice of where to put the 3 heads among the 10 tosses, so there are $\binom{10}{3}$ such arrangements. The probability of obtaining exactly three heads is then

$$p = \frac{\binom{10}{3}}{2^{10}} = 0.1172.$$

(b) Using the same reasoning as in part (a), the number of sequences in the sample space that contain exactly k heads ($k = 0, 1, 2, 3$) is $\binom{10}{k}$. Hence, the probability of obtaining three or fewer heads is

$$p' = \frac{\binom{10}{0} + \binom{10}{1} + \binom{10}{2} + \binom{10}{3}}{2^{10}}$$
$$= \frac{1 + 10 + 45 + 120}{2^{10}} = \frac{176}{2^{10}} = 0.1719.$$ ◂

Note: Using Two Different Methods in the Same Problem. Part (a) of Example 1.8.7 is another example of using two different counting methods in the same problem. Part (b) illustrates another general technique. In this part, we broke the event of interest into several disjoint subsets and counted the numbers of outcomes separately for each subset and then added the counts together to get the total. In many problems, it can require several applications of the same or different counting

methods in order to count the number of outcomes in an event. The next example is one in which the elements of an event are formed in two parts (multiplication rule), but we need to perform separate combination calculations to determine the numbers of outcomes for each part.

Example 1.8.8

Sampling without Replacement. Suppose that a class contains 15 boys and 30 girls, and that 10 students are to be selected at random for a special assignment. We shall determine the probability p that exactly three boys will be selected.

The number of different combinations of the 45 students that might be obtained in the sample of 10 students is $\binom{45}{10}$, and the statement that the 10 students are selected at random means that each of these $\binom{45}{10}$ possible combinations is equally probable. Therefore, we must find the number of these combinations that contain exactly three boys and seven girls.

When a combination of three boys and seven girls is formed, the number of different combinations in which three boys can be selected from the 15 available boys is $\binom{15}{3}$, and the number of different combinations in which seven girls can be selected from the 30 available girls is $\binom{30}{7}$. Since each of these combinations of three boys can be paired with each of the combinations of seven girls to form a distinct sample, the number of combinations containing exactly three boys is $\binom{15}{3}\binom{30}{7}$. Therefore, the desired probability is

$$p = \frac{\binom{15}{3}\binom{30}{7}}{\binom{45}{10}} = 0.2904. \qquad ◂$$

Example 1.8.9

Playing Cards. Suppose that a deck of 52 cards containing four aces is shuffled thoroughly and the cards are then distributed among four players so that each player receives 13 cards. We shall determine the probability that each player will receive one ace.

The number of possible different combinations of the four positions in the deck occupied by the four aces is $\binom{52}{4}$, and it may be assumed that each of these $\binom{52}{4}$ combinations is equally probable. If each player is to receive one ace, then there must be exactly one ace among the 13 cards that the first player will receive and one ace among each of the remaining three groups of 13 cards that the other three players will receive. In other words, there are 13 possible positions for the ace that the first player is to receive, 13 other possible positions for the ace that the second player is to receive, and so on. Therefore, among the $\binom{52}{4}$ possible combinations of the positions for the four aces, exactly 13^4 of these combinations will lead to the desired result. Hence, the probability p that each player will receive one ace is

$$p = \frac{13^4}{\binom{52}{4}} = 0.1055. \qquad ◂$$

Ordered versus Unordered Samples Several of the examples in this section and the previous section involved counting the numbers of possible samples that could arise using various sampling schemes. Sometimes we treated the same collection of elements in different orders as different samples, and sometimes we treated the same elements in different orders as the same sample. In general, how can one tell which is the correct way to count in a given problem? Sometimes, the problem description will make it clear which is needed. For example, if we are asked to find the probability

that the items in a sample arrive in a specified order, then we cannot even specify the event of interest unless we treat different arrangements of the same items as different outcomes. Examples 1.8.5 and 1.8.6 illustrate how different problem descriptions can lead to very different calculations.

However, there are cases in which the problem description does not make it clear whether or not one must count the same elements in different orders as different outcomes. Indeed, there are some problems that can be solved correctly both ways. Example 1.8.9 is one such problem. In that problem, we needed to decide what we would call an outcome, and then we needed to count how many outcomes were in the whole sample space S and how many were in the event E of interest. In the solution presented in Example 1.8.9, we chose as our outcomes the positions in the 52-card deck that were occupied by the four aces. We did not count different arrangements of the four aces in those four positions as different outcomes when we counted the number of outcomes in S. Hence, when we calculated the number of outcomes in E, we also did not count the different arrangements of the four aces in the four possible positions as different outcomes. In general, this is the principle that should guide the choice of counting method. If we have the choice between whether or not to count the same elements in different orders as different outcomes, then we need to make our choice and be consistent throughout the problem. If we count the same elements in different orders as different outcomes when counting the outcomes in S, we must do the same when counting the elements of E. If we do not count them as different outcomes when counting S, we should not count them as different when counting E.

Example 1.8.10

Playing Cards, Revisited. We shall solve the problem in Example 1.8.9 again, but this time, we shall distinguish outcomes with the same cards in different orders. To go to the extreme, let each outcome be a complete ordering of the 52 cards. So, there are 52! possible outcomes. How many of these have one ace in each of the four sets of 13 cards received by the four players? As before, there are 13^4 ways to choose the four positions for the four aces, one among each of the four sets of 13 cards. No matter which of these sets of positions we choose, there are 4! ways to arrange the four aces in these four positions. No matter how the aces are arranged, there are 48! ways to arrange the remaining 48 cards in the 48 remaining positions. So, there are $13^4 \times 4! \times 48!$ outcomes in the event of interest. We then calculate

$$p = \frac{13^4 \times 4! \times 48!}{52!} = 0.1055. \qquad \blacktriangleleft$$

In the following example, whether one counts the same items in different orders as different outcomes is allowed to depend on which events one wishes to use.

Example 1.8.11

Lottery Tickets. In a lottery game, six numbers from 1 to 30 are drawn at random from a bin without replacement, and each player buys a ticket with six different numbers from 1 to 30. If all six numbers drawn match those on the player's ticket, the player wins. We assume that all possible draws are equally likely. One way to construct a sample space for the experiment of drawing the winning combination is to consider the possible sequences of draws. That is, each outcome consists of an ordered subset of six numbers chosen from the 30 available numbers. There are $P_{30,6} = 30!/24!$ such outcomes. With this sample space S, we can calculate probabilities for events such as

$$A = \{\text{the draw contains the numbers 1, 14, 15, 20, 23, and 27}\},$$
$$B = \{\text{one of the numbers drawn is 15}\}, \text{ and}$$
$$C = \{\text{the first number drawn is less than 10}\}.$$

There is another natural sample space, which we shall denote S', for this experiment. It consists solely of the different combinations of six numbers drawn from the 30 available. There are $\binom{30}{6} = 30!/(6!24!)$ such outcomes. It also seems natural to consider all of these outcomes equally likely. With this sample space, we can calculate the probabilities of the events A and B above, but C is not a subset of the sample space S', so we cannot calculate its probability using this smaller sample space. When the sample space for an experiment could naturally be constructed in more than one way, one needs to choose based on for which events one wants to compute probabilities. ◀

Example 1.8.11 raises the question of whether one will compute the same probabilities using two different sample spaces when the event, such as A or B, exists in both sample spaces. In the example, each outcome in the smaller sample space S' corresponds to an event in the larger sample space S. Indeed, each outcome s' in S' corresponds to the event in S containing the 6! permutations of the single combination s'. For example, the event A in the example has only one outcome $s' = (1, 14, 15, 20, 23, 27)$ in the sample space S', while the corresponding event in the sample space S has 6! permutations including

$$(1, 14, 15, 20, 23, 27), \ (14, 20, 27, 15, 23, 1), \ (27, 23, 20, 15, 14, 1), \ \text{etc.}$$

In the sample space S, the probability of the event A is

$$\Pr(A) = \frac{6!}{P_{30,6}} = \frac{6!24!}{30!} = \frac{1}{\binom{30}{6}}.$$

In the sample space S', the event A has this same probability because it has only one of the $\binom{30}{6}$ equally likely outcomes. The same reasoning applies to every outcome in S'. Hence, if the same event can be expressed in both sample spaces S and S', we will compute the same probability using either sample space. This is a special feature of examples like Example 1.8.11 in which each outcome in the smaller sample space corresponds to an event in the larger sample space with the same number of elements. There are examples in which this feature is not present, and one cannot treat both sample spaces as simple sample spaces.

Example 1.8.12

Tossing Coins. An experiment consists of tossing a coin two times. If we want to distinguish H followed by T from T followed by H, we should use the sample space $S = \{HH, HT, TH, TT\}$, which might naturally be assumed a simple sample space. On the other hand, we might be interested solely in the number of H's tossed. In this case, we might consider the smaller sample space $S' = \{0, 1, 2\}$ where each outcome merely counts the number of H's. The outcomes 0 and 2 in S' each correspond to a single outcome in S, but $1 \in S'$ corresponds to the event $\{HT, TH\} \subset S$ with two outcomes. If we think of S as a simple sample space, then S' will not be a simple sample space, because the outcome 1 will have probability 1/2 while the other two outcomes each have probability 1/4.

There are situations in which one would be justified in treating S' as a simple sample space and assigning each of its outcomes probability 1/3. One might do this if one believed that the coin was not fair, but one had no idea how unfair it was or which side were more likely to land up. In this case, S would not be a simple sample space, because two of its outcomes would have probability 1/3 and the other two would have probabilities that add up to 1/3. ◀

Example 1.8.6 is another case of two different sample spaces in which each outcome in one sample space corresponds to a different number of outcomes in the other space. See Exercise 12 in Sec. 1.9 for a more complete analysis of Example 1.8.6.

The Tennis Tournament

We shall now present a difficult problem that has a simple and elegant solution. Suppose that n tennis players are entered in a tournament. In the first round, the players are paired one against another at random. The loser in each pair is eliminated from the tournament, and the winner in each pair continues into the second round. If the number of players n is odd, then one player is chosen at random before the pairings are made for the first round, and that player automatically continues into the second round. All the players in the second round are then paired at random. Again, the loser in each pair is eliminated, and the winner in each pair continues into the third round. If the number of players in the second round is odd, then one of these players is chosen at random before the others are paired, and that player automatically continues into the third round. The tournament continues in this way until only two players remain in the final round. They then play against each other, and the winner of this match is the winner of the tournament. We shall assume that all n players have equal ability, and we shall determine the probability p that two specific players A and B will ever play against each other during the tournament.

We shall first determine the total number of matches that will be played during the tournament. After each match has been played, one player—the loser of that match—is eliminated from the tournament. The tournament ends when everyone has been eliminated from the tournament except the winner of the final match. Since exactly $n-1$ players must be eliminated, it follows that exactly $n-1$ matches must be played during the tournament.

The number of possible pairs of players is $\binom{n}{2}$. Each of the two players in every match is equally likely to win that match, and all initial pairings are made in a random manner. Therefore, before the tournament begins, every possible pair of players is equally likely to appear in each particular one of the $n-1$ matches to be played during the tournament. Accordingly, the probability that players A and B will meet in some particular match that is specified in advance is $1/\binom{n}{2}$. If A and B do meet in that particular match, one of them will lose and be eliminated. Therefore, these same two players cannot meet in more than one match.

It follows from the preceding explanation that the probability p that players A and B will meet at some time during the tournament is equal to the product of the probability $1/\binom{n}{2}$ that they will meet in any particular specified match and the total number $n-1$ of different matches in which they might possibly meet. Hence,

$$p = \frac{n-1}{\binom{n}{2}} = \frac{2}{n}.$$

Summary

We showed that the number of size k subsets of a set of size n is $\binom{n}{k} = n!/[k!(n-k)!]$. This turns out to be the number of possible samples of size k drawn without replacement from a population of size n as well as the number of arrangements of n items of two types with k of one type and $n-k$ of the other type. We also saw several

examples in which more than one counting technique was required at different points in the same problem. Sometimes, more than one technique is required to count the elements of a single set.

Exercises

1. Two pollsters will canvas a neighborhood with 20 houses. Each pollster will visit 10 of the houses. How many different assignments of pollsters to houses are possible?

2. Which of the following two numbers is larger: $\binom{93}{30}$ or $\binom{93}{31}$?

3. Which of the following two numbers is larger: $\binom{93}{30}$ or $\binom{93}{63}$?

4. A box contains 24 light bulbs, of which four are defective. If a person selects four bulbs from the box at random, without replacement, what is the probability that all four bulbs will be defective?

5. Prove that the following number is an integer:

$$\frac{4155 \times 4156 \times \cdots \times 4250 \times 4251}{2 \times 3 \times \cdots \times 96 \times 97}.$$

6. Suppose that n people are seated in a random manner in a row of n theater seats. What is the probability that two particular people A and B will be seated next to each other?

7. If k people are seated in a random manner in a row containing n seats ($n > k$), what is the probability that the people will occupy k adjacent seats in the row?

8. If k people are seated in a random manner in a circle containing n chairs ($n > k$), what is the probability that the people will occupy k adjacent chairs in the circle?

9. If n people are seated in a random manner in a row containing $2n$ seats, what is the probability that no two people will occupy adjacent seats?

10. A box contains 24 light bulbs, of which two are defective. If a person selects 10 bulbs at random, without replacement, what is the probability that both defective bulbs will be selected?

11. Suppose that a committee of 12 people is selected in a random manner from a group of 100 people. Determine the probability that two particular people A and B will both be selected.

12. Suppose that 35 people are divided in a random manner into two teams in such a way that one team contains 10 people and the other team contains 25 people. What is the probability that two particular people A and B will be on the same team?

13. A box contains 24 light bulbs of which four are defective. If one person selects 10 bulbs from the box in a random manner, and a second person then takes the remaining 14 bulbs, what is the probability that all four defective bulbs will be obtained by the same person?

14. Prove that, for all positive integers n and k ($n \geq k$),

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}.$$

15.

a. Prove that

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n.$$

b. Prove that

$$\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \cdots + (-1)^n \binom{n}{n} = 0.$$

Hint: Use the binomial theorem.

16. The United States Senate contains two senators from each of the 50 states. **(a)** If a committee of eight senators is selected at random, what is the probability that it will contain at least one of the two senators from a certain specified state? **(b)** What is the probability that a group of 50 senators selected at random will contain one senator from each state?

17. A deck of 52 cards contains four aces. If the cards are shuffled and distributed in a random manner to four players so that each player receives 13 cards, what is the probability that all four aces will be received by the same player?

18. Suppose that 100 mathematics students are divided into five classes, each containing 20 students, and that awards are to be given to 10 of these students. If each student is equally likely to receive an award, what is the probability that exactly two students in each class will receive awards?

19. A restaurant has n items on its menu. During a particular day, k customers will arrive and each one will choose one item. The manager wants to count how many different collections of customer choices are possible without regard to the order in which the choices are made. (For example, if $k = 3$ and $a_1, \ldots, a_n$ are the menu items,

then $a_1a_3a_1$ is not distinguished from $a_1a_1a_3$.) Prove that the number of different collections of customer choices is $\binom{n+k-1}{k}$. *Hint:* Assume that the menu items are $a_1, \ldots, a_n$. Show that each collection of customer choices, arranged with the a_1's first, the a_2's second, etc., can be identified with a sequence of k zeros and $n-1$ ones, where each 0 stands for a customer choice and each 1 indicates a point in the sequence where the menu item number increases by 1. For example, if $k=3$ and $n=5$, then $a_1a_1a_3$ becomes 0011011.

20. Prove the binomial theorem 1.8.2. *Hint:* You may use an *induction* argument. That is, first prove that the result is true if $n=1$. Then, under the assumption that there is n_0 such that the result is true for all $n \le n_0$, prove that it is also true for $n = n_0 + 1$.

21. Return to the birthday problem on page 30. How many different sets of birthdays are available with k people and 365 days when we don't distinguish the same birthdays in different orders? For example, if $k=3$, we would count (Jan. 1, Mar. 3, Jan.1) the same as (Jan. 1, Jan. 1, Mar. 3).

22. Let n be a large even integer. Use Stirlings' formula (Theorem 1.7.5) to find an approximation to the binomial coefficient $\binom{n}{n/2}$. Compute the approximation with $n = 500$.

1.9 Multinomial Coefficients

We learn how to count the number of ways to partition a finite set into more than two disjoint subsets. This generalizes the binomial coefficients from Sec. 1.8. The generalization is useful when outcomes consist of several parts selected from a fixed number of distinct types.

We begin with a fairly simple example that will illustrate the general ideas of this section.

Example 1.9.1

Choosing Committees. Suppose that 20 members of an organization are to be divided into three committees A, B, and C in such a way that each of the committees A and B is to have eight members and committee C is to have four members. We shall determine the number of different ways in which members can be assigned to these committees. Notice that each of the 20 members gets assigned to one and only one committee.

One way to think of the assignments is to form committee A first by choosing its eight members and then split the remaining 12 members into committees B and C. Each of these operations is choosing a combination, and every choice of committee A can be paired with every one of the splits of the remaining 12 members into committees B and C. Hence, the number of assignments into three committees is the product of the numbers of combinations for the two parts of the assignment. Specifically, to form committee A, we must choose eight out of 20 members, and this can be done in $\binom{20}{8}$ ways. Then to split the remaining 12 members into committees B and C there are are $\binom{12}{8}$ ways to do it. Here, the answer is

$$\binom{20}{8}\binom{12}{8} = \frac{20!}{8!12!}\frac{12!}{8!4!} = \frac{20!}{8!8!4!} = 62{,}355{,}150. \blacktriangleleft$$

Notice how the 12! that appears in the denominator of $\binom{20}{8}$ divides out with the 12! that appears in the numerator of $\binom{12}{8}$. This fact is the key to the general formula that we shall derive next.

In general, suppose that n distinct elements are to be divided into k different groups ($k \ge 2$) in such a way that, for $j = 1, \ldots, k$, the jth group contains exactly n_j elements, where $n_1 + n_2 + \cdots + n_k = n$. It is desired to determine the number of different ways in which the n elements can be divided into the k groups. The

n_1 elements in the first group can be selected from the n available elements in $\binom{n}{n_1}$ different ways. After the n_1 elements in the first group have been selected, the n_2 elements in the second group can be selected from the remaining $n - n_1$ elements in $\binom{n-n_1}{n_2}$ different ways. Hence, the total number of different ways of selecting the elements for both the first group and the second group is $\binom{n}{n_1}\binom{n-n_1}{n_2}$. After the $n_1 + n_2$ elements in the first two groups have been selected, the number of different ways in which the n_3 elements in the third group can be selected is $\binom{n-n_1-n_2}{n_3}$. Hence, the total number of different ways of selecting the elements for the first three groups is

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}.$$

It follows from the preceding explanation that, for each $j = 1, \ldots, k-2$ after the first j groups have been formed, the number of different ways in which the n_{j+1} elements in the next group $(j+1)$ can be selected from the remaining $n - n_1 - \cdots - n_j$ elements is $\binom{n-n_1-\cdots-n_j}{n_{j+1}}$. After the elements of group $k-1$ have been selected, the remaining n_k elements must then form the last group. Hence, the total number of different ways of dividing the n elements into the k groups is

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-\cdots-n_{k-2}}{n_{k-1}} = \frac{n!}{n_1!n_2!\cdots n_k!},$$

where the last formula follows from writing the binomial coefficients in terms of factorials.

Definition 1.9.1 Multinomial Coefficients. The number

$$\frac{n!}{n_1!n_2!\cdots n_k!}, \quad \text{which we shall denote by} \quad \binom{n}{n_1, n_2, \ldots, n_k},$$

is called a *multinomial coefficient*.

The name *multinomial coefficient* derives from the appearance of the symbol in the multinomial theorem, whose proof is left as Exercise 11 in this section.

Theorem 1.9.1 Multinomial Theorem. For all numbers $x_1, \ldots, x_k$ and each positive integer n,

$$(x_1 + \cdots + x_k)^n = \sum \binom{n}{n_1, n_2, \ldots, n_k} x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k},$$

where the summation extends over all possible combinations of nonnegative integers $n_1, \ldots, n_k$ such that $n_1 + n_2 + \cdots + n_k = n$. ■

A multinomial coefficient is a generalization of the binomial coefficient discussed in Sec. 1.8. For $k = 2$, the multinomial theorem is the same as the binomial theorem, and the multinomial coefficient becomes a binomial coefficient. In particular,

$$\binom{n}{k, n-k} = \binom{n}{k}.$$

Example 1.9.2 Choosing Committees. In Example 1.9.1, we see that the solution obtained there is the same as the multinomial coefficient for which $n = 20$, $k = 3$, $n_1 = n_2 = 8$, and $n_3 = 4$, namely,

$$\binom{20}{8, 8, 4} = \frac{20!}{(8!)^2 4!} = 62{,}355{,}150.$$ ◄

Arrangements of Elements of More Than Two Distinct Types Just as binomial coefficients can be used to represent the number of different arrangements of the elements of a set containing elements of only two distinct types, multinomial coefficients can be used to represent the number of different arrangements of the elements of a set containing elements of k different types ($k \geq 2$). Suppose, for example, that n balls of k different colors are to be arranged in a row and that there are n_j balls of color j ($j = 1, \ldots, k$), where $n_1 + n_2 + \cdots + n_k = n$. Then each different arrangement of the n balls corresponds to a different way of dividing the n available positions in the row into a group of n_1 positions to be occupied by the balls of color 1, a second group of n_2 positions to be occupied by the balls of color 2, and so on. Hence, the total number of different possible arrangements of the n balls must be

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}.$$

Example 1.9.3

Rolling Dice. Suppose that 12 dice are to be rolled. We shall determine the probability p that each of the six different numbers will appear twice.

Each outcome in the sample space S can be regarded as an ordered sequence of 12 numbers, where the ith number in the sequence is the outcome of the ith roll. Hence, there will be 6^{12} possible outcomes in S, and each of these outcomes can be regarded as equally probable. The number of these outcomes that would contain each of the six numbers $1, 2, \ldots, 6$ exactly twice will be equal to the number of different possible arrangements of these 12 elements. This number can be determined by evaluating the multinomial coefficient for which $n = 12$, $k = 6$, and $n_1 = n_2 = \cdots = n_6 = 2$. Hence, the number of such outcomes is

$$\binom{12}{2, 2, 2, 2, 2, 2} = \frac{12!}{(2!)^6},$$

and the required probability p is

$$p = \frac{12!}{2^6 6^{12}} = 0.0034.$$ ◄

Example 1.9.4

Playing Cards. A deck of 52 cards contains 13 hearts. Suppose that the cards are shuffled and distributed among four players A, B, C, and D so that each player receives 13 cards. We shall determine the probability p that player A will receive six hearts, player B will receive four hearts, player C will receive two hearts, and player D will receive one heart.

The total number N of different ways in which the 52 cards can be distributed among the four players so that each player receives 13 cards is

$$N = \binom{52}{13, 13, 13, 13} = \frac{52!}{(13!)^4}.$$

It may be assumed that each of these ways is equally probable. We must now calculate the number M of ways of distributing the cards so that each player receives the required number of hearts. The number of different ways in which the hearts can be distributed to players A, B, C, and D so that the numbers of hearts they receive are 6, 4, 2, and 1, respectively, is

$$\binom{13}{6, 4, 2, 1} = \frac{13!}{6!4!2!1!}.$$

Also, the number of different ways in which the other 39 cards can then be distributed to the four players so that each will have a total of 13 cards is

$$\binom{39}{7, 9, 11, 12} = \frac{39!}{7!9!11!12!}.$$

Therefore,

$$M = \frac{13!}{6!4!2!1!} \cdot \frac{39!}{7!9!11!12!},$$

and the required probability p is

$$p = \frac{M}{N} = \frac{13!39!(13!)^4}{6!4!2!1!7!9!11!12!52!} = 0.00196.$$

There is another approach to this problem along the lines indicated in Example 1.8.9 on page 37. The number of possible different combinations of the 13 positions in the deck occupied by the hearts is $\binom{52}{13}$. If player A is to receive six hearts, there are $\binom{13}{6}$ possible combinations of the six positions these hearts occupy among the 13 cards that A will receive. Similarly, if player B is to receive four hearts, there are $\binom{13}{4}$ possible combinations of their positions among the 13 cards that B will receive. There are $\binom{13}{2}$ possible combinations for player C, and there are $\binom{13}{1}$ possible combinations for player D. Hence,

$$p = \frac{\binom{13}{6}\binom{13}{4}\binom{13}{2}\binom{13}{1}}{\binom{52}{13}},$$

which produces the same value as the one obtained by the first method of solution. ◀

Summary

Multinomial coefficients generalize binomial coefficients. The coefficient $\binom{n}{n_1, \ldots, n_k}$ is the number of ways to partition a set of n items into distinguishable subsets of sizes $n_1, \ldots, n_k$ where $n_1 + \cdots + n_k = n$. It is also the number of arrangements of n items of k different types for which n_i are of type i for $i = 1, \ldots, k$. Example 1.9.4 illustrates another important point to remember about computing probabilities: There might be more than one correct method for computing the same probability.

Exercises

1. Three pollsters will canvas a neighborhood with 21 houses. Each pollster will visit seven of the houses. How many different assignments of pollsters to houses are possible?

2. Suppose that 18 red beads, 12 yellow beads, eight blue beads, and 12 black beads are to be strung in a row. How many different arrangements of the colors can be formed?

3. Suppose that two committees are to be formed in an organization that has 300 members. If one committee is to have five members and the other committee is to have eight members, in how many different ways can these committees be selected?

4. If the letters $s, s, s, t, t, t, i, i, a, c$ are arranged in a random order, what is the probability that they will spell the word "statistics"?

5. Suppose that n balanced dice are rolled. Determine the probability that the number j will appear exactly n_j times ($j = 1, \ldots, 6$), where $n_1 + n_2 + \ldots + n_6 = n$.

6. If seven balanced dice are rolled, what is the probability that each of the six different numbers will appear at least once?

7. Suppose that a deck of 25 cards contains 12 red cards. Suppose also that the 25 cards are distributed in a random manner to three players A, B, and C in such a way that player A receives 10 cards, player B receives eight cards, and player C receives seven cards. Determine the probability that player A will receive six red cards, player B will receive two red cards, and player C will receive four red cards.

8. A deck of 52 cards contains 12 picture cards. If the 52 cards are distributed in a random manner among four players in such a way that each player receives 13 cards, what is the probability that each player will receive three picture cards?

9. Suppose that a deck of 52 cards contains 13 red cards, 13 yellow cards, 13 blue cards, and 13 green cards. If the 52 cards are distributed in a random manner among four players in such a way that each player receives 13 cards, what is the probability that each player will receive 13 cards of the same color?

10. Suppose that two boys named Davis, three boys named Jones, and four boys named Smith are seated at random in a row containing nine seats. What is the probability that the Davis boys will occupy the first two seats in the row, the Jones boys will occupy the next three seats, and the Smith boys will occupy the last four seats?

11. Prove the multinomial theorem 1.9.1. (You may wish to use the same hint as in Exercise 20 in Sec. 1.8.)

12. Return to Example 1.8.6. Let S be the larger sample space (first method of choosing) and let S' be the smaller sample space (second method). For each element s' of S', let $N(s')$ stand for the number of elements of S that lead to the same boxful s' when the order of choosing is ignored.

a. For each $s' \in S'$, find a formula for $N(s')$. *Hint:* Let n_i stand for the number of items of type i in s' for $i = 1, \dots, 7$.

b. Verify that $\sum_{s' \in S'} N(s')$ equals the number of outcomes in S.

1.10 The Probability of a Union of Events

The axioms of probability tell us directly how to find the probability of the union of disjoint events. Theorem 1.5.7 showed how to find the probability for the union of two arbitrary events. This theorem is generalized to the union of an arbitrary finite collection of events.

We shall now consider again an arbitrary sample space S that may contain either a finite number of outcomes or an infinite number, and we shall develop some further general properties of the various probabilities that might be specified for the events in S. In this section, we shall study in particular the probability of the union $\bigcup_{i=1}^{n} A_i$ of n events $A_1, \dots, A_n$.

If the events $A_1, \dots, A_n$ are disjoint, we know that

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \Pr(A_i).$$

Furthermore, for every two events A_1 and A_2, regardless of whether or not they are disjoint, we know from Theorem 1.5.7 of Sec. 1.5 that

$$\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2).$$

In this section, we shall extend this result, first to three events and then to an arbitrary finite number of events.

The Union of Three Events

Theorem 1.10.1 For every three events A_1, A_2, and A_3,

$$
\begin{aligned}
\Pr(A_1 \cup A_2 \cup A_3) = {} & \Pr(A_1) + \Pr(A_2) + \Pr(A_3) \\
& - [\Pr(A_1 \cap A_2) + \Pr(A_2 \cap A_3) + \Pr(A_1 \cap A_3)] \\
& + \Pr(A_1 \cap A_2 \cap A_3).
\end{aligned} \tag{1.10.1}
$$

Proof By the associative property of unions (Theorem 1.4.6), we can write

$$A_1 \cup A_2 \cup A_3 = (A_1 \cup A_2) \cup A_3.$$

Apply Theorem 1.5.7 to the two events $A = A_1 \cup A_2$ and $B = A_3$ to obtain

$$
\begin{aligned}
\Pr(A_1 \cup A_2 \cup A_3) = {} & \Pr(A \cup B) \\
& = \Pr(A) + \Pr(B) - \Pr(A \cap B).
\end{aligned} \tag{1.10.2}
$$

We next compute the three probabilities on the far right side of (1.10.2) and combine them to get (1.10.1). First, apply Theorem 1.5.7 to the two events A_1 and A_2 to obtain

$$\Pr(A) = \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2). \tag{1.10.3}$$

Next, use the first distributive property in Theorem 1.4.10 to write

$$A \cap B = (A_1 \cup A_2) \cap A_3 = (A_1 \cap A_3) \cup (A_2 \cap A_3). \tag{1.10.4}$$

Apply Theorem 1.5.7 to the events on the far right side of (1.10.4) to obtain

$$\Pr(A \cap B) = \Pr(A_1 \cap A_3) + \Pr(A_2 \cap A_3) - \Pr(A_1 \cap A_2 \cap A_3). \tag{1.10.5}$$

Substitute (1.10.3), $\Pr(B) = \Pr(A_3)$, and (1.10.5) into (1.10.2) to complete the proof. ■

Example 1.10.1 Student Enrollment. Among a group of 200 students, 137 students are enrolled in a mathematics class, 50 students are enrolled in a history class, and 124 students are enrolled in a music class. Furthermore, the number of students enrolled in both the mathematics and history classes is 33, the number enrolled in both the history and music classes is 29, and the number enrolled in both the mathematics and music classes is 92. Finally, the number of students enrolled in all three classes is 18. We shall determine the probability that a student selected at random from the group of 200 students will be enrolled in at least one of the three classes.

Let A_1 denote the event that the selected student is enrolled in the mathematics class, let A_2 denote the event that he is enrolled in the history class, and let A_3 denote the event that he is enrolled in the music class. To solve the problem, we must determine the value of $\Pr(A_1 \cup A_2 \cup A_3)$. From the given numbers,

$$
\begin{aligned}
& \Pr(A_1) = \frac{137}{200}, \quad \Pr(A_2) = \frac{50}{200}, \quad \Pr(A_3) = \frac{124}{200}, \\
& \Pr(A_1 \cap A_2) = \frac{33}{200}, \quad \Pr(A_2 \cap A_3) = \frac{29}{200}, \quad \Pr(A_1 \cap A_3) = \frac{92}{200}, \\
& \Pr(A_1 \cap A_2 \cap A_3) = \frac{18}{200}.
\end{aligned}
$$

It follows from Eq. (1.10.1) that $\Pr(A_1 \cup A_2 \cup A_3) = 175/200 = 7/8$. ◄

The Union of a Finite Number of Events

A result similar to Theorem 1.10.1 holds for any arbitrary finite number of events, as shown by the following theorem.

Theorem 1.10.2 For every n events $A_1, \ldots, A_n$,

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \Pr(A_i) - \sum_{i<j} \Pr(A_i \cap A_j) + \sum_{i<j<k} \Pr(A_i \cap A_j \cap A_k) - \sum_{i<j<k<l} \Pr(A_i \cap A_j \cap A_k \cap A_l) + \cdots + (-1)^{n+1} \Pr(A_1 \cap A_2 \cap \cdots \cap A_n). \quad (1.10.6)$$

Proof The proof proceeds by induction. In particular, we first establish that (1.10.6) is true for $n = 1$ and $n = 2$. Next, we show that if there exists m such that (1.10.6) is true for all $n \le m$, then (1.10.6) is also true for $n = m + 1$. The case of $n = 1$ is trivial, and the case of $n = 2$ is Theorem 1.5.7. To complete the proof, assume that (1.10.6) is true for all $n \le m$. Let $A_1, \ldots, A_{m+1}$ be events. Define $A = \bigcup_{i=1}^{m} A_i$ and $B = A_{m+1}$. Theorem 1.5.7 says that

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \quad (1.10.7)$$

We have assumed that $\Pr(A)$ equals (1.10.6) with $n = m$. We need to show that when we add $\Pr(A)$ to $\Pr(B) - \Pr(A \cap B)$, we get (1.10.6) with $n = m + 1$. The difference between (1.10.6) with $n = m + 1$ and $\Pr(A)$ is all of the terms in which one of the subscripts (i, j, k, etc.) equals $m + 1$. Those terms are the following:

$$\Pr(A_{m+1}) - \sum_{i=1}^{m} \Pr(A_i \cap A_{m+1}) + \sum_{i<j} \Pr(A_i \cap A_j \cap A_{m+1}) - \sum_{i<j<k} \Pr(A_i \cap A_j \cap A_k \cap A_{m+1}) + \cdots + (-1)^{m+2} \Pr(A_1 \cap A_2 \cap \cdots \cap A_m \cap A_{m+1}). \quad (1.10.8)$$

The first term in (1.10.8) is $\Pr(B) = \Pr(A_{m+1})$. All that remains is to show that $-\Pr(A \cap B)$ equals all but the first term in (1.10.8).

Use the natural generalization of the distributive property (Theorem 1.4.10) to write

$$A \cap B = \left(\bigcup_{i=1}^{m} A_i\right) \cap A_{m+1} = \bigcup_{i=1}^{m} (A_i \cap A_{m+1}). \quad (1.10.9)$$

The union in (1.10.9) contains m events, and hence we can apply (1.10.6) with $n = m$ and each A_i replaced by $A_i \cap A_{m+1}$. The result is that $-\Pr(A \cap B)$ equals all but the first term in (1.10.8). ■

The calculation in Theorem 1.10.2 can be outlined as follows: First, take the sum of the probabilities of the n individual events. Second, subtract the sum of the probabilities of the intersections of all possible pairs of events; in this step, there will be $\binom{n}{2}$ different pairs for which the probabilities are included. Third, add the probabilities of the intersections of all possible groups of three of the events; there will be $\binom{n}{3}$ intersections of this type. Fourth, subtract the sum of the probabilities of the intersections of all possible groups of four of the events; there will be $\binom{n}{4}$ intersections of this type. Continue in this way until, finally, the probability of the intersection of all n events is either added or subtracted, depending on whether n is an odd number or an even number.

The Matching Problem

Suppose that all the cards in a deck of n different cards are placed in a row, and that the cards in another similar deck are then shuffled and placed in a row on top of the cards in the original deck. It is desired to determine the probability p_n that there will be at least one match between the corresponding cards from the two decks. The same problem can be expressed in various entertaining contexts. For example, we could suppose that a person types n letters, types the corresponding addresses on n envelopes, and then places the n letters in the n envelopes in a random manner. It could be desired to determine the probability p_n that at least one letter will be placed in the correct envelope. As another example, we could suppose that the photographs of n famous film actors are paired in a random manner with n photographs of the same actors taken when they were babies. It could then be desired to determine the probability p_n that the photograph of at least one actor will be paired correctly with this actor's own baby photograph.

Here we shall discuss this matching problem in the context of letters being placed in envelopes. Thus, we shall let A_i be the event that letter i is placed in the correct envelope $(i = 1, \ldots, n)$, and we shall determine the value of $p_n = \Pr\left(\bigcup_{i=1}^{n} A_i\right)$ by using Eq. (1.10.6). Since the letters are placed in the envelopes at random, the probability $\Pr(A_i)$ that any particular letter will be placed in the correct envelope is $1/n$. Therefore, the value of the first summation on the right side of Eq. (1.10.6) is

$$\sum_{i=1}^{n} \Pr(A_i) = n \cdot \frac{1}{n} = 1.$$

Furthermore, since letter 1 could be placed in any one of n envelopes and letter 2 could then be placed in any one of the other $n - 1$ envelopes, the probability $\Pr(A_1 \cap A_2)$ that both letter 1 and letter 2 will be placed in the correct envelopes is $1/[n(n-1)]$. Similarly, the probability $\Pr(A_i \cap A_j)$ that any two specific letters i and j $(i \neq j)$ will both be placed in the correct envelopes is $1/[n(n-1)]$. Therefore, the value of the second summation on the right side of Eq. (1.10.6) is

$$\sum_{i<j} \Pr(A_i \cap A_j) = \binom{n}{2} \frac{1}{n(n-1)} = \frac{1}{2!}.$$

By similar reasoning, it can be determined that the probability $\Pr(A_i \cap A_j \cap A_k)$ that any three specific letters i, j, and k $(i < j < k)$ will be placed in the correct envelopes is $1/[n(n-1)(n-2)]$. Therefore, the value of the third summation is

$$\sum_{i<j<k} \Pr(A_i \cap A_j \cap A_k) = \binom{n}{3} \frac{1}{n(n-1)(n-2)} = \frac{1}{3!}.$$

This procedure can be continued until it is found that the probability $\Pr(A_1 \cap A_2 \cdots \cap A_n)$ that all n letters will be placed in the correct envelopes is $1/(n!)$. It now follows from Eq. (1.10.6) that the probability p_n that at least one letter will be placed in the correct envelope is

$$p_n = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + (-1)^{n+1}\frac{1}{n!}. \tag{1.10.10}$$

This probability has the following interesting features. As $n \to \infty$, the value of p_n approaches the following limit:

$$\lim_{n\to\infty} p_n = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots.$$

It is shown in books on elementary calculus that the sum of the infinite series on the right side of this equation is $1 - (1/e)$, where $e = 2.71828\ldots$. Hence, $1 - (1/e) = 0.63212\ldots$. It follows that for a large value of n, the probability p_n that at least one letter will be placed in the correct envelope is approximately 0.63212.

The exact values of p_n, as given in Eq. (1.10.10), will form an oscillating sequence as n increases. As n increases through the even integers $2, 4, 6, \ldots$, the values of p_n will increase toward the limiting value 0.63212; and as n increases through the odd integers $3, 5, 7, \ldots$, the values of p_n will decrease toward this same limiting value.

The values of p_n converge to the limit very rapidly. In fact, for $n = 7$ the exact value p_7 and the limiting value of p_n agree to four decimal places. Hence, regardless of whether seven letters are placed at random in seven envelopes or seven million letters are placed at random in seven million envelopes, the probability that at least one letter will be placed in the correct envelope is 0.6321.

Summary

We generalized the formula for the probability of the union of two arbitrary events to the union of finitely many events. As an aside, there are cases in which it is easier to compute $\Pr(A_1 \cup \ldots \cup A_n)$ as $1 - \Pr(A_1^c \cap \cdots \cap A_n^c)$ using the fact that $(A_1 \cup \ldots \cup A_n)^c = A_1^c \cap \cdots \cap A_n^c$.

Exercises

1. Three players are each dealt, in a random manner, five cards from a deck containing 52 cards. Four of the 52 cards are aces. Find the probability that at least one person receives exactly two aces in their five cards.

2. In a certain city, three newspapers A, B, and C are published. Suppose that 60 percent of the families in the city subscribe to newspaper A, 40 percent of the families subscribe to newspaper B, and 30 percent subscribe to newspaper C. Suppose also that 20 percent of the families subscribe to both A and B, 10 percent subscribe to both A and C, 20 percent subscribe to both B and C, and 5 percent subscribe to all three newspapers A, B, and C. What percentage of the families in the city subscribe to at least one of the three newspapers?

3. For the conditions of Exercise 2, what percentage of the families in the city subscribe to exactly one of the three newspapers?

4. Suppose that three compact discs are removed from their cases, and that after they have been played, they are put back into the three empty cases in a random manner. Determine the probability that at least one of the CD's will be put back into the proper cases.

5. Suppose that four guests check their hats when they arrive at a restaurant, and that these hats are returned to them in a random order when they leave. Determine the probability that no guest will receive the proper hat.

6. A box contains 30 red balls, 30 white balls, and 30 blue balls. If 10 balls are selected at random, without replacement, what is the probability that at least one color will be missing from the selection?

7. Suppose that a school band contains 10 students from the freshman class, 20 students from the sophomore class, 30 students from the junior class, and 40 students from the senior class. If 15 students are selected at random from the band, what is the probability that at least one student will be selected from each of the four classes? *Hint:* First determine the probability that at least one of the four classes will not be represented in the selection.

8. If n letters are placed at random in n envelopes, what is the probability that exactly $n - 1$ letters will be placed in the correct envelopes?

9. Suppose that n letters are placed at random in n envelopes, and let q_n denote the probability that no letter is placed in the correct envelope. For which of the following four values of n is q_n largest: $n = 10$, $n = 21$, $n = 53$, or $n = 300$?

10. If three letters are placed at random in three envelopes, what is the probability that exactly one letter will be placed in the correct envelope?

11. Suppose that 10 cards, of which five are red and five are green, are placed at random in 10 envelopes, of which five are red and five are green. Determine the probability that exactly x envelopes will contain a card with a matching color ($x = 0, 1, \ldots, 10$).

12. Let $A_1, A_2, \ldots$ be an infinite sequence of events such that $A_1 \subset A_2 \subset \cdots$. Prove that

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} \Pr(A_n).$$

Hint: Let the sequence $B_1, B_2, \ldots$ be defined as in Exercise 12 of Sec. 1.5, and show that

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} \Pr\left(\bigcup_{i=1}^{n} B_i\right) = \lim_{n\to\infty} \Pr(A_n).$$

13. Let $A_1, A_2, \ldots$ be an infinite sequence of events such that $A_1 \supset A_2 \supset \cdots$. Prove that

$$\Pr\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} \Pr(A_n).$$

Hint: Consider the sequence $A_1^c, A_2^c, \ldots$, and apply Exercise 12.

1.11 Statistical Swindles

This section presents some examples of how one can be misled by arguments that require one to ignore the calculus of probability.

Misleading Use of Statistics

The field of statistics has a poor image in the minds of many people because there is a widespread belief that statistical data and statistical analyses can easily be manipulated in an unscientific and unethical fashion in an effort to show that a particular conclusion or point of view is correct. We all have heard the sayings that "There are three kinds of lies: lies, damned lies, and statistics" (Mark Twain [1924, p. 246] says that this line has been attributed to Benjamin Disraeli) and that "you can prove anything with statistics."

One benefit of studying probability and statistics is that the knowledge we gain enables us to analyze statistical arguments that we read in newspapers, magazines, or elsewhere. We can then evaluate these arguments on their merits, rather than accepting them blindly. In this section, we shall describe three schemes that have been used to induce consumers to send money to the operators of the schemes in exchange for certain types of information. The first two schemes are not strictly statistical in nature, but they are strongly based on undertones of probability.

Perfect Forecasts

Suppose that one Monday morning you receive in the mail a letter from a firm with which you are not familiar, stating that the firm sells forecasts about the stock market for very high fees. To indicate the firm's ability in forecasting, it predicts that a particular stock, or a particular portfolio of stocks, will rise in value during the coming week. You do not respond to this letter, but you do watch the stock market during the week and notice that the prediction was correct. On the following Monday morning you receive another letter from the same firm containing another prediction, this one specifying that a particular stock will drop in value during the coming week. Again the prediction proves to be correct.

This routine continues for seven weeks. Every Monday morning you receive a prediction in the mail from the firm, and each of these seven predictions proves to be correct. On the eighth Monday morning, you receive another letter from the firm. This letter states that for a large fee the firm will provide another prediction, on the basis of which you can presumably make a large amount of money on the stock market. How should you respond to this letter?

Since the firm has made seven successive correct predictions, it would seem that it must have some special information about the stock market and is not simply guessing. After all, the probability of correctly guessing the outcomes of seven successive tosses of a fair coin is only $(1/2)^7 = 0.008$. Hence, if the firm had only been guessing each week, then the firm had a probability less than 0.01 of being correct seven weeks in a row.

The fallacy here is that you may have seen only a relatively small number of the forecasts that the firm made during the seven-week period. Suppose, for example, that the firm started the entire process with a list of $2^7 = 128$ potential clients. On the first Monday, the firm could send the forecast that a particular stock will rise in value to half of these clients and send the forecast that the same stock will drop in value to the other half. On the second Monday, the firm could continue writing to those 64 clients for whom the first forecast proved to be correct. It could again send a new forecast to half of those 64 clients and the opposite forecast to the other half. At the end of seven weeks, the firm (which usually consists of only one person and a computer) must necessarily have one client (and only one client) for whom all seven forecasts were correct.

By following this procedure with several different groups of 128 clients, and starting new groups each week, the firm may be able to generate enough positive responses from clients for it to realize significant profits.

Guaranteed Winners

There is another scheme that is somewhat related to the one just described but that is even more elegant because of its simplicity. In this scheme, a firm advertises that for a fixed fee, usually 10 or 20 dollars, it will send the client its forecast of the winner of any upcoming baseball game, football game, boxing match, or other sports event that the client might specify. Furthermore, the firm offers a money-back guarantee that this forecast will be correct; that is, if the team or person designated as the winner in the forecast does not actually turn out to be the winner, the firm will return the full fee to the client.

How should you react to such an advertisement? At first glance, it would appear that the firm must have some special knowledge about these sports events, because otherwise it could not afford to guarantee its forecasts. Further reflection reveals, however, that the firm simply cannot lose, because its only expenses are those for advertising and postage. In effect, when this scheme is used, the firm holds the client's fee until the winner has been decided. If the forecast was correct, the firm keeps the fee; otherwise, it simply returns the fee to the client.

On the other hand, the client can very well lose. He presumably purchases the firm's forecast because he desires to bet on the sports event. If the forecast proves to be wrong, the client will not have to pay any fee to the firm, but he will have lost any money that he bet on the predicted winner.

Thus, when there are "guaranteed winners," only the firm is guaranteed to win. In fact, the firm knows that it will be able to keep the fees from all the clients for whom the forecasts were correct.

Improving Your Lottery Chances

State lotteries have become very popular in America. People spend millions of dollars each week to purchase tickets with very small chances of winning medium to enormous prizes. With so much money being spent on lottery tickets, it should not be surprising that a few enterprising individuals have concocted schemes to cash in on the probabilistic naïveté of the ticket-buying public. There are now several books and videos available that claim to help lottery players improve their performance. People actually pay money for these items. Some of the advice is just common sense, but some of it is misleading and plays on subtle misconceptions about probability.

For concreteness, suppose that we have a game in which there are 40 balls numbered 1 to 40 and six are drawn without replacement to determine the winning combination. A ticket purchase requires the customer to choose six different numbers from 1 to 40 and pay a fee. This game has $\binom{40}{6} = 3{,}838{,}380$ different winning combinations and the same number of possible tickets. One piece of advice often found in published lottery aids is not to choose the six numbers on your ticket too far apart. Many people tend to pick their six numbers uniformly spread out from 1 to 40, but the winning combination often has two consecutive numbers or at least two numbers very close together. Some of these "advisors" recommend that, since it is more likely that there will be numbers close together, players should bunch some of their six numbers close together. Such advice might make sense in order to avoid choosing the same numbers as other players in a parimutuel game (i.e., a game in which all winners share the jackpot). But the idea that any strategy can improve your chances of winning is misleading.

To see why this advice is misleading, let E be the event that the winning combination contains at least one pair of consecutive numbers. The reader can calculate $\Pr(E)$ in Exercise 13 in Sec. 1.12. For this example, $\Pr(E) = 0.577$. So the lottery aids are correct that E has high probability. However, by claiming that choosing a ticket in E increases your chance of winning, they confuse the probability of the event E with the probability of each outcome in E. If you choose the ticket (5, 7, 14, 23, 24, 38), your probability of winning is only 1/3,828,380, just as it would be if you chose any other ticket. The fact that this ticket happens to be in E doesn't make your probability of winning equal to 0.577. The reason that $\Pr(E)$ is so big is that so many different combinations are in E. Each of those combinations still has probability 1/3,828,380 of winning, and you only get one combination on each ticket. The fact that there are so many combinations in E does not make each one any more likely than anything else.

1.12 Supplementary Exercises

1. Suppose that a coin is tossed seven times. Let A denote the event that a head is obtained on the first toss, and let B denote the event that a head is obtained on the fifth toss. Are A and B disjoint?

2. If A, B, and D are three events such that $\Pr(A \cup B \cup D) = 0.7$, what is the value of $\Pr(A^c \cap B^c \cap D^c)$?

3. Suppose that a certain precinct contains 350 voters, of which 250 are Democrats and 100 are Republicans. If 30 voters are chosen at random from the precinct, what is the probability that exactly 18 Democrats will be selected?

4. Suppose that in a deck of 20 cards, each card has one of the numbers 1, 2, 3, 4, or 5 and there are four cards with each number. If 10 cards are chosen from the deck at random, without replacement, what is the probability that each of the numbers 1, 2, 3, 4, and 5 will appear exactly twice?

5. Consider the contractor in Example 1.5.4 on page 19. He wishes to compute the probability that the total utility demand is high, meaning that the sum of water and electrical demand (in the units of Example 1.4.5) is at least

215. Draw a picture of this event on a graph like Fig. 1.5 or Fig. 1.9 and find its probability.

6. Suppose that a box contains r red balls and w white balls. Suppose also that balls are drawn from the box one at a time, at random, without replacement. **(a)** What is the probability that all r red balls will be obtained before any white balls are obtained? **(b)** What is the probability that all r red balls will be obtained before two white balls are obtained?

7. Suppose that a box contains r red balls, w white balls, and b blue balls. Suppose also that balls are drawn from the box one at a time, at random, without replacement. What is the probability that all r red balls will be obtained before any white balls are obtained?

8. Suppose that 10 cards, of which seven are red and three are green, are put at random into 10 envelopes, of which seven are red and three are green, so that each envelope contains one card. Determine the probability that exactly k envelopes will contain a card with a matching color ($k = 0, 1, \dots, 10$).

9. Suppose that 10 cards, of which five are red and five are green, are put at random into 10 envelopes, of which seven are red and three are green, so that each envelope contains one card. Determine the probability that exactly k envelopes will contain a card with a matching color ($k = 0, 1, \dots, 10$).

10. Suppose that the events A and B are disjoint. Under what conditions are A^c and B^c disjoint?

11. Let A_1, A_2, and A_3 be three arbitrary events. Show that the probability that exactly one of these three events will occur is

$$\begin{aligned}&\Pr(A_1) + \Pr(A_2) + \Pr(A_3)\\ &\qquad - 2\Pr(A_1 \cap A_2) - 2\Pr(A_1 \cap A_3) - 2\Pr(A_2 \cap A_3)\\ &\qquad + 3\Pr(A_1 \cap A_2 \cap A_3).\end{aligned}$$

12. Let $A_1, \dots, A_n$ be n arbitrary events. Show that the probability that exactly one of these n events will occur is

$$\sum_{i=1}^{n} \Pr(A_i) - 2\sum_{i<j} \Pr(A_i \cap A_j) + 3 \sum_{i<j<k} \Pr(A_i \cap A_j \cap A_k)$$
$$- \cdots + (-1)^{n+1} n \Pr(A_1 \cap A_2 \cdots \cap A_n).$$

13. Consider a state lottery game in which each winning combination and each ticket consists of one set of k numbers chosen from the numbers 1 to n without replacement. We shall compute the probability that the winning combination contains at least one pair of consecutive numbers.

a. Prove that if $n < 2k - 1$, then every winning combination has at least one pair of consecutive numbers. For the rest of the problem, assume that $n \geq 2k - 1$.

b. Let $i_1 < \cdots < i_k$ be an arbitrary possible winning combination arranged in order from smallest to largest. For $s = 1, \dots, k$, let $j_s = i_s - (s - 1)$. That is,

$$\begin{aligned} j_1 &= i_1,\\ j_2 &= i_2 - 1\\ &\vdots\\ j_k &= i_k - (k - 1).\end{aligned}$$

Prove that $(i_1, \dots, i_k)$ contains at least one pair of consecutive numbers if and only if $(j_1, \dots, j_k)$ contains repeated numbers.

c. Prove that $1 \leq j_1 \leq \cdots \leq j_k \leq n - k + 1$ and that the number of $(j_1, \dots, j_k)$ sets with no repeats is $\binom{n-k+1}{k}$.

d. Find the probability that there is no pair of consecutive numbers in the winning combination.

e. Find the probability of at least one pair of consecutive numbers in the winning combination.

Chapter 2

Conditional Probability

2.1 The Definition of Conditional Probability

A major use of probability in statistical inference is the updating of probabilities when certain events are observed. The updated probability of event A after we learn that event B has occurred is the conditional probability of A given B.

Example 2.1.1 Lottery Ticket. Consider a state lottery game in which six numbers are drawn without replacement from a bin containing the numbers 1–30. Each player tries to match the set of six numbers that will be drawn without regard to the order in which the numbers are drawn. Suppose that you hold a ticket in such a lottery with the numbers 1, 14, 15, 20, 23, and 27. You turn on your television to watch the drawing but all you see is one number, 15, being drawn when the power suddenly goes off in your house. You don't even know whether 15 was the first, last, or some in-between draw. However, now that you know that 15 appears in the winning draw, the probability that your ticket is a winner must be higher than it was before you saw the draw. How do you calculate the revised probability? ◂

Example 2.1.1 is typical of the following situation. An experiment is performed for which the sample space S is given (or can be constructed easily) and the probabilities are available for all of the events of interest. We then learn that some event B has occured, and we want to know how the probability of another event A changes after we learn that B has occurred. In Example 2.1.1, the event that we have learned is $B = \{$one of the numbers drawn is 15$\}$. We are certainly interested in the probability of

$$A = \{\text{the numbers 1, 14, 15, 20, 23, and 27 are drawn}\},$$

and possibly other events.

If we know that the event B has occurred, then we know that the outcome of the experiment is one of those included in B. Hence, to evaluate the probability that A will occur, we must consider the set of those outcomes in B that also result in the occurrence of A. As sketched in Fig. 2.1, this set is precisely the set $A \cap B$. It is therefore natural to calculate the revised probability of A according to the following definition.

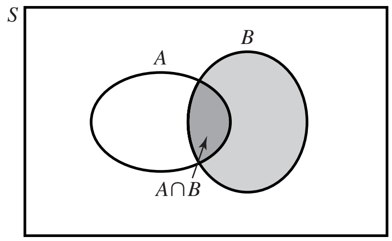


Figure 2.1 The outcomes in the event B that also belong to the event A.

Definition 2.1.1 Conditional Probability. Suppose that we learn that an event B has occurred and that we wish to compute the probability of another event A taking into account that we know that B has occurred. The new probability of A is called the *conditional probability of the event A given that the event B has occurred* and is denoted $\Pr(A|B)$. If $\Pr(B) > 0$, we compute this probability as

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}. \tag{2.1.1}$$

The conditional probability $\Pr(A|B)$ is not defined if $\Pr(B) = 0$.

For convenience, the notation in Definition 2.1.1 is read simply as the conditional probability of A given B. Eq. (2.1.1) indicates that $\Pr(A|B)$ is computed as the proportion of the total probability $\Pr(B)$ that is represented by $\Pr(A \cap B)$, intuitively the proportion of B that is also part of A.

Example 2.1.2 Lottery Ticket. In Example 2.1.1, you learned that the event

$$B = \{\text{one of the numbers drawn is 15}\}$$

has occurred. You want to calculate the probability of the event A that your ticket is a winner. Both events A and B are expressible in the sample space that consists of the $\binom{30}{6} = 30!/(6!24!)$ possible combinations of 30 items taken six at a time, namely, the unordered draws of six numbers from 1–30. The event B consists of combinations that include 15. Since there are 29 remaining numbers from which to choose the other five in the winning draw, there are $\binom{29}{5}$ outcomes in B. It follows that

$$\Pr(B) = \frac{\binom{29}{5}}{\binom{30}{6}} = \frac{29!24!6!}{30!5!24!} = 0.2.$$

The event A that your ticket is a winner consists of a single outcome that is also in B, so $A \cap B = A$, and

$$\Pr(A \cap B) = \Pr(A) = \frac{1}{\binom{30}{6}} = \frac{6!24!}{30!} = 1.68 \times 10^{-6}.$$

It follows that the conditional probability of A given B is

$$\Pr(A|B) = \frac{\frac{6!24!}{30!}}{0.2} = 8.4 \times 10^{-6}.$$

This is five times as large as $\Pr(A)$ before you learned that B had occurred. ◀

Definition 2.1.1 for the conditional probability $\Pr(A|B)$ is worded in terms of the subjective interpretation of probability in Sec. 1.2. Eq. (2.1.1) also has a simple meaning in terms of the frequency interpretation of probability. According to the

frequency interpretation, if an experimental process is repeated a large number of times, then the proportion of repetitions in which the event B will occur is approximately $\Pr(B)$ and the proportion of repetitions in which both the event A and the event B will occur is approximately $\Pr(A \cap B)$. Therefore, among those repetitions in which the event B occurs, the proportion of repetitions in which the event A will also occur is approximately equal to

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Example 2.1.3 Rolling Dice. Suppose that two dice were rolled and it was observed that the sum T of the two numbers was odd. We shall determine the probability that T was less than 8.

If we let A be the event that $T < 8$ and let B be the event that T is odd, then $A \cap B$ is the event that T is 3, 5, or 7. From the probabilities for two dice given at the end of Sec. 1.6, we can evaluate $\Pr(A \cap B)$ and $\Pr(B)$ as follows:

$$\Pr(A \cap B) = \frac{2}{36} + \frac{4}{36} + \frac{6}{36} = \frac{12}{36} = \frac{1}{3},$$

$$\Pr(B) = \frac{2}{36} + \frac{4}{36} + \frac{6}{36} + \frac{4}{36} + \frac{2}{36} = \frac{18}{36} = \frac{1}{2}.$$

Hence,

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{2}{3}.$$ ◂

Example 2.1.4 A Clinical Trial. It is very common for patients with episodes of depression to have a recurrence within two to three years. Prien et al. (1984) studied three treatments for depression: imipramine, lithium carbonate, and a combination. As is traditional in such studies (called *clinical trials*), there was also a group of patients who received a placebo. (A placebo is a treatment that is supposed to be neither helpful nor harmful. Some patients are given a placebo so that they will not know that they did not receive one of the other treatments. None of the other patients knew which treatment or placebo they received, either.) In this example, we shall consider 150 patients who entered the study after an episode of depression that was classified as "unipolar" (meaning that there was no manic disorder). They were divided into the four groups (three treatments plus placebo) and followed to see how many had recurrences of depression. Table 2.1 summarizes the results. If a patient were selected at random from this study and it were found that the patient received the placebo treatment, what is the conditional probability that the patient had a relapse? Let B be the event that the patient received the placebo, and let A be the event that

Table 2.1 Results of the clinical depression study in Example 2.1.4

	Treatment group				
Response	Imipramine	Lithium	Combination	Placebo	*Total*
Relapse	18	13	22	24	77
No relapse	22	25	16	10	73
Total	40	38	38	34	150

the patient had a relapse. We can calculate $\Pr(B) = 34/150$ and $\Pr(A \cap B) = 24/150$ directly from the table. Then $\Pr(A|B) = 24/34 = 0.706$. On the other hand, if the randomly selected patient is found to have received lithium (call this event C) then $\Pr(C) = 38/150$, $\Pr(A \cap C) = 13/150$, and $\Pr(A|C) = 13/38 = 0.342$. Knowing which treatment a patient received seems to make a difference to the probability of relapse. In Chapter 10, we shall study methods for being more precise about how much of a difference it makes. ◀

Example 2.1.5 **Rolling Dice Repeatedly.** Suppose that two dice are to be rolled repeatedly and the sum T of the two numbers is to be observed for each roll. We shall determine the probability p that the value $T = 7$ will be observed before the value $T = 8$ is observed.

The desired probability p could be calculated directly as follows: We could assume that the sample space S contains all sequences of outcomes that terminate as soon as either the sum $T = 7$ or the sum $T = 8$ is obtained. Then we could find the sum of the probabilities of all the sequences that terminate when the value $T = 7$ is obtained.

However, there is a simpler approach in this example. We can consider the simple experiment in which two dice are rolled. If we repeat the experiment until either the sum $T = 7$ or the sum $T = 8$ is obtained, the effect is to restrict the outcome of the experiment to one of these two values. Hence, the problem can be restated as follows: Given that the outcome of the experiment is either $T = 7$ or $T = 8$, determine the probability p that the outcome is actually $T = 7$.

If we let A be the event that $T = 7$ and let B be the event that the value of T is either 7 or 8, then $A \cap B = A$ and

$$p = \Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A)}{\Pr(B)}.$$

From the probabilities for two dice given in Example 1.6.5, $\Pr(A) = 6/36$ and $\Pr(B) = (6/36) + (5/36) = 11/36$. Hence, $p = 6/11$. ◀

The Multiplication Rule for Conditional Probabilities

In some experiments, certain conditional probabilities are relatively easy to assign directly. In these experiments, it is then possible to compute the probability that both of two events occur by applying the next result that follows directly from Eq. (2.1.1) and the analogous definition of $\Pr(B|A)$.

Theorem 2.1.1 **Multiplication Rule for Conditional Probabilities.** Let A and B be events. If $\Pr(B) > 0$, then

$$\Pr(A \cap B) = \Pr(B)\Pr(A|B).$$

If $\Pr(A) > 0$, then

$$\Pr(A \cap B) = \Pr(A)\Pr(B|A).$$ ■

Example 2.1.6 **Selecting Two Balls.** Suppose that two balls are to be selected at random, without replacement, from a box containing r red balls and b blue balls. We shall determine the probability p that the first ball will be red and the second ball will be blue.

Let A be the event that the first ball is red, and let B be the event that the second ball is blue. Obviously, $\Pr(A) = r/(r + b)$. Furthermore, if the event A has occurred, then one red ball has been removed from the box on the first draw. Therefore, the

probability of obtaining a blue ball on the second draw will be

$$\Pr(B|A) = \frac{b}{r+b-1}.$$

It follows that

$$\Pr(A \cap B) = \frac{r}{r+b} \cdot \frac{b}{r+b-1}. \blacktriangleleft$$

The principle that has just been applied can be extended to any finite number of events, as stated in the following theorem.

Theorem 2.1.2 Multiplication Rule for Conditional Probabilities. Suppose that $A_1, A_2, \ldots, A_n$ are events such that $\Pr(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$. Then

$$\Pr(A_1 \cap A_2 \cap \cdots \cap A_n) = \Pr(A_1)\Pr(A_2|A_1)\Pr(A_3|A_1 \cap A_2)\cdots\Pr(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}). \tag{2.1.2}$$

Proof The product of probabilities on the right side of Eq. (2.1.2) is equal to

$$\Pr(A_1) \cdot \frac{\Pr(A_1 \cap A_2)}{\Pr(A_1)} \cdot \frac{\Pr(A_1 \cap A_2 \cap A_3)}{\Pr(A_1 \cap A_2)} \cdots \frac{\Pr(A_1 \cap A_2 \cap \cdots \cap A_n)}{\Pr(A_1 \cap A_2 \cdots \cap A_{n-1})}.$$

Since $\Pr(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$, each of the denominators in this product must be positive. All of the terms in the product cancel each other except the final numerator $\Pr(A_1 \cap A_2 \cap \cdots \cap A_n)$, which is the left side of Eq. (2.1.2). ■

Example 2.1.7 Selecting Four Balls. Suppose that four balls are selected one at a time, without replacement, from a box containing r red balls and b blue balls ($r \geq 2$, $b \geq 2$). We shall determine the probability of obtaining the sequence of outcomes red, blue, red, blue.

If we let R_j denote the event that a red ball is obtained on the jth draw and let B_j denote the event that a blue ball is obtained on the jth draw ($j = 1, \ldots, 4$), then

$$\begin{aligned}\Pr(R_1 \cap B_2 \cap R_3 \cap B_4) &= \Pr(R_1)\Pr(B_2|R_1)\Pr(R_3|R_1 \cap B_2)\Pr(B_4|R_1 \cap B_2 \cap R_3) \\ &= \frac{r}{r+b} \cdot \frac{b}{r+b-1} \cdot \frac{r-1}{r+b-2} \cdot \frac{b-1}{r+b-3}. \blacktriangleleft\end{aligned}$$

Note: Conditional Probabilities Behave Just Like Probabilities. In all of the situations that we shall encounter in this text, every result that we can prove has a conditional version given an event B with $\Pr(B) > 0$. Just replace *all* probabilities by conditional probabilities given B and replace all conditional probabilities given other events C by conditional probabilities given $C \cap B$. For example, Theorem 1.5.3 says that $\Pr(A^c) = 1 - \Pr(A)$. It is easy to prove that $\Pr(A^c|B) = 1 - \Pr(A|B)$ if $\Pr(B) > 0$. (See Exercises 11 and 12 in this section.) Another example is Theorem 2.1.3, which is a conditional version of the multiplication rule Theorem 2.1.2. Although a proof is given for Theorem 2.1.3, we shall not provide proofs of all such conditional theorems, because their proofs are generally very similar to the proofs of the unconditional versions.

Theorem 2.1.3 Suppose that $A_1, A_2, \ldots, A_n, B$ are events such that $\Pr(B) > 0$ and $\Pr(A_1 \cap A_2 \cap \cdots \cap A_{n-1}|B) > 0$. Then

$$\begin{aligned}\Pr(A_1 \cap A_2 \cap \cdots \cap A_n|B) = \Pr(A_1|B)\Pr(A_2|A_1 \cap B)\cdots \\ \times \Pr(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1} \cap B).\end{aligned} \tag{2.1.3}$$

Proof The product of probabilities on the right side of Eq. (2.1.3) is equal to

$$\frac{\Pr(A_1 \cap B)}{\Pr(B)} \cdot \frac{\Pr(A_1 \cap A_2 \cap B)}{\Pr(A_1 \cap B)} \cdots \frac{\Pr(A_1 \cap A_2 \cap \cdots \cap A_n B)}{\Pr(A_1 \cap A_2 \cdots \cap A_{n-1} \cap B)}.$$

Since $\Pr(A_1 \cap A_2 \cap \cdots \cap A_{n-1}|B) > 0$, each of the denominators in this product must be positive. All of the terms in the product cancel each other except the first denominator and the final numerator to yield $\Pr(A_1 \cap A_2 \cap \cdots \cap A_n \cap B)/\Pr(B)$, which is the left side of Eq. (2.1.3). ■

Conditional Probability and Partitions

Theorem 1.4.11 shows how to calculate the probability of an event by partitioning the sample space into two events B and B^c. This result easily generalizes to larger partitions, and when combined with Theorem 2.1.1 it leads to a very powerful tool for calculating probabilities.

Definition 2.1.2 Partition. Let S denote the sample space of some experiment, and consider k events $B_1, \ldots, B_k$ in S such that $B_1, \ldots, B_k$ are disjoint and $\bigcup_{i=1}^{k} B_i = S$. It is said that these events form a *partition* of S.

Typically, the events that make up a partition are chosen so that an important source of uncertainty in the problem is reduced if we learn which event has occurred.

Example 2.1.8 Selecting Bolts. Two boxes contain long bolts and short bolts. Suppose that one box contains 60 long bolts and 40 short bolts, and that the other box contains 10 long bolts and 20 short bolts. Suppose also that one box is selected at random and a bolt is then selected at random from that box. We would like to determine the probability that this bolt is long. ◄

Partitions can facilitate the calculations of probabilities of certain events.

Theorem 2.1.4 Law of total probability. Suppose that the events $B_1, \ldots, B_k$ form a partition of the space S and $\Pr(B_j) > 0$ for $j = 1, \ldots, k$. Then, for every event A in S,

$$\Pr(A) = \sum_{j=1}^{k} \Pr(B_j)\Pr(A|B_j). \tag{2.1.4}$$

Proof The events $B_1 \cap A, B_2 \cap A, \ldots, B_k \cap A$ will form a partition of A, as illustrated in Fig. 2.2. Hence, we can write

$$A = (B_1 \cap A) \cup (B_2 \cap A) \cup \cdots \cup (B_k \cap A).$$

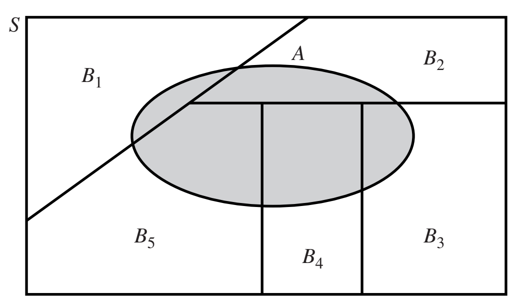


Figure 2.2 The intersections of A with events $B_1, \ldots, B_5$ of a partition in the proof of Theorem 2.1.4.

Furthermore, since the k events on the right side of this equation are disjoint,

$$\Pr(A) = \sum_{j=1}^{k} \Pr(B_j \cap A).$$

Finally, if $\Pr(B_j) > 0$ for $j = 1, \ldots, k$, then $\Pr(B_j \cap A) = \Pr(B_j)\Pr(A|B_j)$ and it follows that Eq. (2.1.4) holds. ■

Example 2.1.9

Selecting Bolts. In Example 2.1.8, let B_1 be the event that the first box (the one with 60 long and 40 short bolts) is selected, let B_2 be the event that the second box (the one with 10 long and 20 short bolts) is selected, and let A be the event that a long bolt is selected. Then

$$\Pr(A) = \Pr(B_1)\Pr(A|B_1) + \Pr(B_2)\Pr(A|B_2).$$

Since a box is selected at random, we know that $\Pr(B_1) = \Pr(B_2) = 1/2$. Furthermore, the probability of selecting a long bolt from the first box is $\Pr(A|B_1) = 60/100 = 3/5$, and the probability of selecting a long bolt from the second box is $\Pr(A|B_2) = 10/30 = 1/3$. Hence,

$$\Pr(A) = \frac{1}{2}\cdot\frac{3}{5} + \frac{1}{2}\cdot\frac{1}{3} = \frac{7}{15}. \quad ◀$$

Example 2.1.10

Achieving a High Score. Suppose that a person plays a game in which his score must be one of the 50 numbers 1, 2, . . . , 50 and that each of these 50 numbers is equally likely to be his score. The first time he plays the game, his score is X. He then continues to play the game until he obtains another score Y such that $Y \geq X$. We will assume that, conditional on previous plays, the 50 scores remain equally likely on all subsequent plays. We shall determine the probability of the event A that $Y = 50$.

For each $i = 1, \ldots, 50$, let B_i be the event that $X = i$. Conditional on B_i, the value of Y is equally likely to be any one of the numbers $i, i+1, \ldots, 50$. Since each of these $(51 - i)$ possible values for Y is equally likely, it follows that

$$\Pr(A|B_i) = \Pr(Y = 50|B_i) = \frac{1}{51 - i}.$$

Furthermore, since the probability of each of the 50 values of X is 1/50, it follows that $\Pr(B_i) = 1/50$ for all i and

$$\Pr(A) = \sum_{i=1}^{50} \frac{1}{50}\cdot\frac{1}{51-i} = \frac{1}{50}\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{50}\right) = 0.0900. \quad ◀$$

Note: Conditional Version of Law of Total Probability. The law of total probability has an analog conditional on another event C, namely,

$$\Pr(A|C) = \sum_{j=1}^{k} \Pr(B_j|C)\Pr(A|B_j \cap C). \tag{2.1.5}$$

The reader can prove this in Exercise 17.

Augmented Experiment In some experiments, it may not be clear from the initial description of the experiment that a partition exists that will facilitate the calculation of probabilities. However, there are many such experiments in which such a partition exists if we imagine that the experiment has some additional structure. Consider the following modification of Examples 2.1.8 and 2.1.9.

Example 2.1.11 Selecting Bolts. There is one box of bolts that contains some long and some short bolts. A manager is unable to open the box at present, so she asks her employees what is the composition of the box. One employee says that it contains 60 long bolts and 40 short bolts. Another says that it contains 10 long bolts and 20 short bolts. Unable to reconcile these opinions, the manager decides that each of the employees is correct with probability 1/2. Let B_1 be the event that the box contains 60 long and 40 short bolts, and let B_2 be the event that the box contains 10 long and 20 short bolts. The probability that the first bolt selected is long is now calculated precisely as in Example 2.1.9. ◀

In Example 2.1.11, there is only one box of bolts, but we believe that it has one of two possible compositions. We let the events B_1 and B_2 determine the possible compositions. This type of situation is very common in experiments.

Example 2.1.12 A Clinical Trial. Consider a clinical trial such as the study of treatments for depression in Example 2.1.4. As in many such trials, each patient has two possible outcomes, in this case relapse and no relapse. We shall refer to relapse as “failure” and no relapse as “success.” For now, we shall consider only patients in the imipramine treatment group. If we knew the effectiveness of imipramine, that is, the proportion p of successes among all patients who might receive the treatment, then we might model the patients in our study as having probability p of success. Unfortunately, we do not know p at the start of the trial. In analogy to the box of bolts with unknown composition in Example 2.1.11, we can imagine that the collection of all available patients (from which the 40 imipramine patients in this trial were selected) has two or more possible compositions. We can imagine that the composition of the collection of patients determines the proportion that will be success. For simplicity, in this example, we imagine that there are 11 different possible compositions of the collection of patients. In particular, we assume that the proportions of success for the 11 possible compositions are 0, 1/10, . . . , 9/10, 1. (We shall be able to handle more realistic models for p in Chapter 3.) For example, if we knew that our patients were drawn from a collection with the proportion 3/10 of successes, we would be comfortable saying that the patients in our sample each have success probability $p = 3/10$. The value of p is an important source of uncertainty in this problem, and we shall partition the sample space by the possible values of p. For $j = 1, \dots, 11$, let B_j be the event that our sample was drawn from a collection with proportion $(j - 1)/10$ of successes. We can also identify B_j as the event $\{p = (j - 1)/10\}$.

Now, let E_1 be the event that the first patient in the imipramine group has a success. We defined each event B_j so that $\Pr(E_1|B_j) = (j - 1)/10$. Supppose that, prior to starting the trial, we believe that $\Pr(B_j) = 1/11$ for each j. It follows that

$$\Pr(E_1) = \sum_{j=1}^{11} \frac{1}{11}\frac{j-1}{10} = \frac{55}{110} = \frac{1}{2}, \tag{2.1.6}$$

where the second equality uses the fact that $\sum_{j=1}^{n} j = n(n + 1)/2$. ◀

The events $B_1, B_2, \dots, B_{11}$ in Example 2.1.12 can be thought of in much the same way as the two events B_1 and B_2 that determine the mixture of long and short bolts in Example 2.1.11. There is only one box of bolts, but there is uncertainty about its composition. Similarly in Example 2.1.12, there is only one group of patients, but we believe that it has one of 11 possible compositions determined by the events $B_1, B_2, \dots, B_{11}$. To call these events, they must be subsets of the sample space for the experiment in question. That will be the case in Example 2.1.12 if we imagine that

the experiment consists not only of observing the numbers of successes and failures among the patients but also of potentially observing enough additional patients to be able to compute p, possibly at some time very far in the future. Similarly, in Example 2.1.11, the two events B_1 and B_2 are subsets of the sample space if we imagine that the experiment consists not only of observing one sample bolt but also of potentially observing the entire composition of the box.

Throughout the remainder of this text, we shall implicitly assume that experiments are augmented to include outcomes that determine the values of quantities such as p. We shall not require that we ever get to observe the complete outcome of the experiment so as to tell us precisely what p is, but merely that there is an experiment that includes all of the events of interest to us, including those that determine quantities like p.

Definition 2.1.3 Augmented Experiment. If desired, any experiment can be augmented to include the potential or hypothetical observation of as much additional information as we would find useful to help us calculate any probabilities that we desire.

Definition 2.1.3 is worded somewhat vaguely because it is intended to cover a wide variety of cases. Here is an explicit application to Example 2.1.12.

Example 2.1.13 A Clinical Trial. In Example 2.1.12, we could explicitly assume that there exists an infinite sequence of patients who could be treated with imipramine even though we will observe only finitely many of them. We could let the sample space consist of infinite sequences of the two symbols S and F such as $(S, S, F, S, F, F, F, \ldots)$. Here S in coordinate i means that the ith patient is a success, and F stands for failure. So, the event E_1 in Example 2.1.12 is the event that the first coordinate is S. The example sequence above is then in the event E_1. To accommodate our interpretation of p as the proportion of successes, we can assume that, for every such sequence, the proportion of S's among the first n coordinates gets close to one of the numbers $0, 1/10, \ldots, 9/10, 1$ as n increases. In this way, p is explicitly the limit of the proportion of successes we would observe if we could find a way to observe indefinitely. In Example 2.1.12, B_2 is the event consisting of all the outcomes in which the limit of the proportion of S's equals $1/10$, B_3 is the set of outcomes in which the limit is $2/10$, etc. Also, we observe only the first 40 coordinates of the infinite sequence, but we still behave as if p exists and could be determined if only we could observe forever. ◀

In the remainder of the text, there will be many experiments that we assume are augmented. In such cases, we will mention which quantities (such as p in Example 2.1.13) would be determined by the augmented part of the experiment even if we do not explicitly mention that the experiment is augmented.

The Game of Craps

We shall conclude this section by discussing a popular gambling game called craps. One version of this game is played as follows: A player rolls two dice, and the sum of the two numbers that appear is observed. If the sum on the first roll is 7 or 11, the player wins the game immediately. If the sum on the first roll is 2, 3, or 12, the player loses the game immediately. If the sum on the first roll is 4, 5, 6, 8, 9, or 10, then the two dice are rolled again and again until the sum is either 7 or the original value. If the original value is obtained a second time before 7 is obtained, then the

player wins. If the sum 7 is obtained before the original value is obtained a second time, then the player loses.

We shall now compute the probability $\Pr(W)$, where W is the event that the player will win. Let the sample space S consist of all possible sequences of sums from the rolls of dice that might occur in a game. For example, some of the elements of S are (4, 7), (11), (4, 3, 4), (12), (10, 8, 2, 12, 6, 7), etc. We see that $(11) \in W$ but $(4, 7) \in W^c$, etc.. We begin by noticing that whether or not an outcome is in W depends in a crucial way on the first roll. For this reason, it makes sense to partition W according to the sum on the first roll. Let B_i be the event that the first roll is i for $i = 2, \dots, 12$.

Theorem 2.1.4 tells us that $\Pr(W) = \sum_{i=2}^{12} \Pr(B_i)\Pr(W|B_i)$. Since $\Pr(B_i)$ for each i was computed in Example 1.6.5, we need to determine $\Pr(W|B_i)$ for each i. We begin with $i = 2$. Because the player loses if the first roll is 2, we have $\Pr(W|B_2) = 0$. Similarly, $\Pr(W|B_3) = 0 = \Pr(W|B_{12})$. Also, $\Pr(W|B_7) = 1$ because the player wins if the first roll is 7. Similarly, $\Pr(W|B_{11}) = 1$.

For each first roll $i \in \{4, 5, 6, 8, 9, 10\}$, $\Pr(W|B_i)$ is the probability that, in a sequence of dice rolls, the sum i will be obtained before the sum 7 is obtained. As described in Example 2.1.5, this probability is the same as the probability of obtaining the sum i when the sum must be either i or 7. Hence,

$$\Pr(W|B_i) = \frac{\Pr(B_i)}{\Pr(B_i \cup B_7)}.$$

We compute the necessary values here:

$$\Pr(W|B_4) = \frac{\frac{3}{36}}{\frac{3}{36} + \frac{6}{36}} = \frac{1}{3}, \qquad P(W|B_5) = \frac{\frac{4}{36}}{\frac{4}{36} + \frac{6}{36}} = \frac{2}{5},$$

$$\Pr(W|B_6) = \frac{\frac{5}{36}}{\frac{5}{36} + \frac{6}{36}} = \frac{5}{11}, \qquad \Pr(W|B_8) = \frac{\frac{5}{36}}{\frac{5}{36} + \frac{6}{36}} = \frac{5}{11},$$

$$\Pr(W|B_9) = \frac{\frac{4}{36}}{\frac{4}{36} + \frac{6}{36}} = \frac{2}{5}, \qquad \Pr(W|B_{10}) = \frac{\frac{3}{36}}{\frac{3}{36} + \frac{6}{36}} = \frac{1}{3}.$$

Finally, we compute the sum $\sum_{i=2}^{12} \Pr(B_i)\Pr(W|B_i)$:

$$\Pr(W) = \sum_{i=2}^{12} \Pr(B_i)\Pr(W|B_i) = 0 + 0 + \frac{3}{36}\frac{1}{3} + \frac{4}{36}\frac{2}{5} + \frac{5}{36}\frac{5}{11} + \frac{6}{36}$$
$$+ \frac{5}{36}\frac{5}{11} + \frac{4}{36}\frac{2}{5} + \frac{3}{36}\frac{1}{3} + \frac{2}{36} + 0 = \frac{2928}{5940} = 0.493.$$

Thus, the probability of winning in the game of craps is slightly less than 1/2.

Summary

The revised probability of an event A after learning that event B (with $\Pr(B) > 0$) has occurred is the conditional probability of A given B, denoted by $\Pr(A|B)$ and computed as $\Pr(A \cap B)/\Pr(B)$. Often it is easy to assess a conditional probability, such as $\Pr(A|B)$, directly. In such a case, we can use the multiplication rule for conditional probabilities to compute $\Pr(A \cap B) = \Pr(B)\Pr(A|B)$. All probability results have versions conditional on an event B with $\Pr(B) > 0$: Just change *all* probabilities so that they are conditional on B in addition to anything else they were already

conditional on. For example, the multiplication rule for conditional probabilities becomes $\Pr(A_1 \cap A_2|B) = \Pr(A_1|B)\Pr(A_2|A_1 \cap B)$. A partition is a collection of disjoint events whose union is the whole sample space. To be most useful, a partition is chosen so that an important source of uncertainty is reduced if we learn which one of the partition events occurs. If the conditional probability of an event A is available given each event in a partition, the law of total probability tells how to combine these conditional probabilities to get $\Pr(A)$.

Exercises

1. If $A \subset B$ with $\Pr(B) > 0$, what is the value of $\Pr(A|B)$?

2. If A and B are disjoint events and $\Pr(B) > 0$, what is the value of $\Pr(A|B)$?

3. If S is the sample space of an experiment and A is any event in that space, what is the value of $\Pr(A|S)$?

4. Each time a shopper purchases a tube of toothpaste, he chooses either brand A or brand B. Suppose that for each purchase after the first, the probability is 1/3 that he will choose the same brand that he chose on his preceding purchase and the probability is 2/3 that he will switch brands. If he is equally likely to choose either brand A or brand B on his first purchase, what is the probability that both his first and second purchases will be brand A and both his third and fourth purchases will be brand B?

5. A box contains r red balls and b blue balls. One ball is selected at random and its color is observed. The ball is then returned to the box and k additional balls of the same color are also put into the box. A second ball is then selected at random, its color is observed, and it is returned to the box together with k additional balls of the same color. Each time another ball is selected, the process is repeated. If four balls are selected, what is the probability that the first three balls will be red and the fourth ball will be blue?

6. A box contains three cards. One card is red on both sides, one card is green on both sides, and one card is red on one side and green on the other. One card is selected from the box at random, and the color on one side is observed. If this side is green, what is the probability that the other side of the card is also green?

7. Consider again the conditions of Exercise 2 of Sec. 1.10. If a family selected at random from the city subscribes to newspaper A, what is the probability that the family also subscribes to newspaper B?

8. Consider again the conditions of Exercise 2 of Sec. 1.10. If a family selected at random from the city subscribes to at least one of the three newspapers A, B, and C, what is the probability that the family subscribes to newspaper A?

9. Suppose that a box contains one blue card and four red cards, which are labeled A, B, C, and D. Suppose also that two of these five cards are selected at random, without replacement.

- **a.** If it is known that card A has been selected, what is the probability that both cards are red?
- **b.** If it is known that at least one red card has been selected, what is the probability that both cards are red?

10. Consider the following version of the game of craps: The player rolls two dice. If the sum on the first roll is 7 or 11, the player wins the game immediately. If the sum on the first roll is 2, 3, or 12, the player loses the game immediately. However, if the sum on the first roll is 4, 5, 6, 8, 9, or 10, then the two dice are rolled again and again until the sum is either 7 or 11 or the original value. If the original value is obtained a second time before either 7 or 11 is obtained, then the player wins. If either 7 or 11 is obtained before the original value is obtained a second time, then the player loses. Determine the probability that the player will win this game.

11. For any two events A and B with $\Pr(B) > 0$, prove that $\Pr(A^c|B) = 1 - \Pr(A|B)$.

12. For any three events A, B, and D, such that $\Pr(D) > 0$, prove that $\Pr(A \cup B|D) = \Pr(A|D) + \Pr(B|D) - \Pr(A \cap B|D)$.

13. A box contains three coins with a head on each side, four coins with a tail on each side, and two fair coins. If one of these nine coins is selected at random and tossed once, what is the probability that a head will be obtained?

14. A machine produces defective parts with three different probabilities depending on its state of repair. If the machine is in good working order, it produces defective parts with probability 0.02. If it is wearing down, it produces defective parts with probability 0.1. If it needs maintenance, it produces defective parts with probability 0.3. The probability that the machine is in good working order is 0.8, the probability that it is wearing down is 0.1, and the probability that it needs maintenance is 0.1. Compute the probability that a randomly selected part will be defective.

15. The percentages of voters classed as Liberals in three different election districts are divided as follows: in the first district, 21 percent; in the second district, 45 percent; and in the third district, 75 percent. If a district is selected at random and a voter is selected at random from that district, what is the probability that she will be a Liberal?

16. Consider again the shopper described in Exercise 4. On each purchase, the probability that he will choose the same brand of toothpaste that he chose on his preceding purchase is 1/3, and the probability that he will switch brands is 2/3. Suppose that on his first purchase the probability that he will choose brand A is 1/4 and the probability that he will choose brand B is 3/4. What is the probability that his second purchase will be brand B?

17. Prove the conditional version of the law of total probability (2.1.5).

2.2 Independent Events

If learning that B has occurred does not change the probability of A, then we say that A and B are independent. There are many cases in which events A and B are not independent, but they would be independent if we learned that some other event C had occurred. In this case, A and B are conditionally independent given C.

Example 2.2.1

Tossing Coins. Suppose that a fair coin is tossed twice. The experiment has four outcomes, HH, HT, TH, and TT, that tell us how the coin landed on each of the two tosses. We can assume that this sample space is simple so that each outcome has probability 1/4. Suppose that we are interested in the second toss. In particular, we want to calculate the probability of the event $A = \{\text{H on second toss}\}$. We see that $A = \{\text{HH,TH}\}$, so that $\Pr(A) = 2/4 = 1/2$. If we learn that the first coin landed T, we might wish to compute the conditional probability $\Pr(A|B)$ where $B = \{\text{T on first toss}\}$. Using the definition of conditional probability, we easily compute

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{1/4}{1/2} = \frac{1}{2},$$

because $A \cap B = \{TH\}$ has probability 1/4. We see that $\Pr(A|B) = \Pr(A)$; hence, we don't change the probability of A even after we learn that B has occurred. ◀

Definition of Independence

The conditional probability of the event A given that the event B has occurred is the revised probability of A after we learn that B has occurred. It might be the case, however, that no revision is necessary to the probability of A even after we learn that B occurs. This is precisely what happened in Example 2.2.1. In this case, we say that A and B are *independent events*. As another example, if we toss a coin and then roll a die, we could let A be the event that the die shows 3 and let B be the event that the coin lands with heads up. If the tossing of the coin is done in isolation of the rolling of the die, we might be quite comfortable assigning $\Pr(A|B) = \Pr(A) = 1/6$. In this case, we say that A and B are independent events.

In general, if $\Pr(B) > 0$, the equation $\Pr(A|B) = \Pr(A)$ can be rewritten as $\Pr(A \cap B)/\Pr(B) = \Pr(A)$. If we multiply both sides of this last equation by $\Pr(B)$, we obtain the equation $\Pr(A \cap B) = \Pr(A)\Pr(B)$. In order to avoid the condition $\Pr(B) > 0$, the mathematical definition of the independence of two events is stated as follows:

Definition 2.2.1

Independent Events. Two events A and B are *independent* if

$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

Suppose that $\Pr(A) > 0$ and $\Pr(B) > 0$. Then it follows easily from the definitions of independence and conditional probability that A and B are independent if and only if $\Pr(A|B) = \Pr(A)$ and $\Pr(B|A) = \Pr(B)$.

Independence of Two Events

If two events A and B are considered to be independent because the events are physically unrelated, and if the probabilities $\Pr(A)$ and $\Pr(B)$ are known, then the definition can be used to assign a value to $\Pr(A \cap B)$.

Example 2.2.2

Machine Operation. Suppose that two machines 1 and 2 in a factory are operated independently of each other. Let A be the event that machine 1 will become inoperative during a given 8-hour period, let B be the event that machine 2 will become inoperative during the same period, and suppose that $\Pr(A) = 1/3$ and $\Pr(B) = 1/4$. We shall determine the probability that at least one of the machines will become inoperative during the given period.

The probability $\Pr(A \cap B)$ that both machines will become inoperative during the period is

$$\Pr(A \cap B) = \Pr(A)\Pr(B) = \left(\frac{1}{3}\right)\left(\frac{1}{4}\right) = \frac{1}{12}.$$

Therefore, the probability $\Pr(A \cup B)$ that at least one of the machines will become inoperative during the period is

$$\begin{aligned}\Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B)\\ &= \frac{1}{3} + \frac{1}{4} - \frac{1}{12} = \frac{1}{2}.\end{aligned}$$ ◀

The next example shows that two events A and B, which are physically related, can, nevertheless, satisfy the definition of independence.

Example 2.2.3

Rolling a Die. Suppose that a balanced die is rolled. Let A be the event that an even number is obtained, and let B be the event that one of the numbers 1, 2, 3, or 4 is obtained. We shall show that the events A and B are independent.

In this example, $\Pr(A) = 1/2$ and $\Pr(B) = 2/3$. Furthermore, since $A \cap B$ is the event that either the number 2 or the number 4 is obtained, $\Pr(A \cap B) = 1/3$. Hence, $\Pr(A \cap B) = \Pr(A)\Pr(B)$. It follows that the events A and B are independent events, even though the occurrence of each event depends on the same roll of a die. ◀

The independence of the events A and B in Example 2.2.3 can also be interpreted as follows: Suppose that a person must bet on whether the number obtained on the die will be even or odd, that is, on whether or not the event A will occur. Since three of the possible outcomes of the roll are even and the other three are odd, the person will typically have no preference between betting on an even number and betting on an odd number.

Suppose also that after the die has been rolled, but before the person has learned the outcome and before she has decided whether to bet on an even outcome or on an odd outcome, she is informed that the actual outcome was one of the numbers 1, 2, 3, or 4, i.e., that the event B has occurred. The person now knows that the outcome was 1, 2, 3, or 4. However, since two of these numbers are even and two are odd, the person will typically still have no preference between betting on an even number and betting on an odd number. In other words, the information that the event B has

occurred is of no help to the person who is trying to decide whether or not the event A has occurred.

Independence of Complements In the foregoing discussion of independent events, we stated that if A and B are independent, then the occurrence or nonoccurrence of A should not be related to the occurrence or nonoccurrence of B. Hence, if A and B satisfy the mathematical definition of independent events, then it should also be true that A and B^c are independent events, that A^c and B are independent events, and that A^c and B^c are independent events. One of these results is established in the next theorem.

Theorem 2.2.1 If two events A and B are independent, then the events A and B^c are also independent.

Proof Theorem 1.5.6 says that

$$\Pr(A \cap B^c) = \Pr(A) - \Pr(A \cap B).$$

Furthermore, since A and B are independent events, $\Pr(A \cap B) = \Pr(A)\Pr(B)$. It now follows that

$$\begin{aligned}\Pr(A \cap B^c) &= \Pr(A) - \Pr(A)\Pr(B) = \Pr(A)[1 - \Pr(B)]\\ &= \Pr(A)\Pr(B^c).\end{aligned}$$

Therefore, the events A and B^c are independent. ■

The proof of the analogous result for the events A^c and B is similar, and the proof for the events A^c and B^c is required in Exercise 2 at the end of this section.

Independence of Several Events

The definition of independent events can be extended to any number of events, $A_1, \dots, A_k$. Intuitively, if learning that some of these events do or do not occur does not change our probabilities for any events that depend only on the remaining events, we would say that all k events are independent. The mathematical definition is the following analog to Definition 2.2.1.

Definition 2.2.2 (Mutually) Independent Events. The k events $A_1, \dots, A_k$ are *independent* (or *mutually independent*) if, for every subset $A_{i_1}, \dots, A_{i_j}$ of j of these events ($j = 2, 3, \dots, k$),

$$\Pr(A_{i_1} \cap \dots \cap A_{i_j}) = \Pr(A_{i_1}) \cdots \Pr(A_{i_j}).$$

As an example, in order for three events A, B, and C to be independent, the following four relations must be satisfied:

$$\begin{aligned}\Pr(A \cap B) &= \Pr(A)\Pr(B),\\ \Pr(A \cap C) &= \Pr(A)\Pr(C),\\ \Pr(B \cap C) &= \Pr(B)\Pr(C),\end{aligned} \tag{2.2.1}$$

and

$$\Pr(A \cap B \cap C) = \Pr(A)\Pr(B)\Pr(C). \tag{2.2.2}$$

It is possible that Eq. (2.2.2) will be satisfied, but one or more of the three relations (2.2.1) will not be satisfied. On the other hand, as is shown in the next example,

it is also possible that each of the three relations (2.2.1) will be satisfied but Eq. (2.2.2) will not be satisfied.

Example 2.2.4

Pairwise Independence. Suppose that a fair coin is tossed twice so that the sample space $S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$ is simple. Define the following three events:

$$A = \{\text{H on first toss}\} = \{\text{HH}, \text{HT}\},$$
$$B = \{\text{H on second toss}\} = \{\text{HH}, \text{TH}\}, \text{ and}$$
$$C = \{\text{Both tosses the same}\} = \{\text{HH}, \text{TT}\}.$$

Then $A \cap B = A \cap C = B \cap C = A \cap B \cap C = \{HH\}$. Hence,

$$\Pr(A) = \Pr(B) = \Pr(C) = 1/2$$

and

$$\Pr(A \cap B) = \Pr(A \cap C) = \Pr(B \cap C) = \Pr(A \cap B \cap C) = 1/4.$$

It follows that each of the three relations of Eq. (2.2.1) is satisfied but Eq. (2.2.2) is not satisfied. These results can be summarized by saying that the events A, B, and C are *pairwise independent*, but all three events are not independent. ◀

We shall now present some examples that will illustrate the power and scope of the concept of independence in the solution of probability problems.

Example 2.2.5

Inspecting Items. Suppose that a machine produces a defective item with probability p $(0 < p < 1)$ and produces a nondefective item with probability $1 - p$. Suppose further that six items produced by the machine are selected at random and inspected, and that the results (defective or nondefective) for these six items are independent. We shall determine the probability that exactly two of the six items are defective.

It can be assumed that the sample space S contains all possible arrangements of six items, each one of which might be either defective or nondefective. For $j = 1, \ldots, 6$, we shall let D_j denote the event that the jth item in the sample is defective so that D_j^c is the event that this item is nondefective. Since the outcomes for the six different items are independent, the probability of obtaining any particular sequence of defective and nondefective items will simply be the product of the individual probabilities for the items. For example,

$$\begin{aligned}\Pr(D_1^c \cap D_2 \cap D_3^c \cap D_4^c \cap D_5 \cap D_6^c) &= \Pr(D_1^c)\Pr(D_2)\Pr(D_3^c)\Pr(D_4^c)\Pr(D_5)\Pr(D_6^c)\\ &= (1-p)p(1-p)(1-p)p(1-p) = p^2(1-p)^4.\end{aligned}$$

It can be seen that the probability of any other particular sequence in S containing two defective items and four nondefective items will also be $p^2(1-p)^4$. Hence, the probability that there will be exactly two defectives in the sample of six items can be found by multiplying the probability $p^2(1-p)^4$ of any particular sequence containing two defectives by the possible number of such sequences. Since there are $\binom{6}{2}$ distinct arrangements of two defective items and four nondefective items, the probability of obtaining exactly two defectives is $\binom{6}{2}p^2(1-p)^4$. ◀

Example 2.2.6

Obtaining a Defective Item. For the conditions of Example 2.2.5, we shall now determine the probability that at least one of the six items in the sample will be defective.

Since the outcomes for the different items are independent, the probability that all six items will be nondefective is $(1-p)^6$. Therefore, the probability that at least one item will be defective is $1 - (1-p)^6$. ◀

Example 2.2.7

Tossing a Coin Until a Head Appears. Suppose that a fair coin is tossed until a head appears for the first time, and assume that the outcomes of the tosses are independent. We shall determine the probability p_n that exactly n tosses will be required.

The desired probability is equal to the probability of obtaining $n-1$ tails in succession and then obtaining a head on the next toss. Since the outcomes of the tosses are independent, the probability of this particular sequence of n outcomes is $p_n = (1/2)^n$.

The probability that a head will be obtained sooner or later (or, equivalently, that tails will not be obtained forever) is

$$\sum_{n=1}^{\infty} p_n = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = 1.$$

Since the sum of the probabilities p_n is 1, it follows that the probability of obtaining an infinite sequence of tails without ever obtaining a head must be 0. ◄

Example 2.2.8

Inspecting Items One at a Time. Consider again a machine that produces a defective item with probability p and produces a nondefective item with probability $1-p$. Suppose that items produced by the machine are selected at random and inspected one at a time until exactly five defective items have been obtained. We shall determine the probability p_n that exactly n items ($n \geq 5$) must be selected to obtain the five defectives.

The fifth defective item will be the nth item that is inspected if and only if there are exactly four defectives among the first $n-1$ items and then the nth item is defective. By reasoning similar to that given in Example 2.2.5, it can be shown that the probability of obtaining exactly four defectives and $n-5$ nondefectives among the first $n-1$ items is $\binom{n-1}{4} p^4(1-p)^{n-5}$. The probability that the nth item will be defective is p. Since the first event refers to outcomes for only the first $n-1$ items and the second event refers to the outcome for only the nth item, these two events are independent. Therefore, the probability that both events will occur is equal to the product of their probabilities. It follows that

$$p_n = \binom{n-1}{4} p^5(1-p)^{n-5}.$$ ◄

Example 2.2.9

People v. Collins. Finkelstein and Levin (1990) describe a criminal case whose verdict was overturned by the Supreme Court of California in part due to a probability calculation involving both conditional probability and independence. The case, *People v. Collins*, 68 Cal. 2d 319, 438 P.2d 33 (1968), involved a purse snatching in which witnesses claimed to see a young woman with blond hair in a ponytail fleeing from the scene in a yellow car driven by a black man with a beard. A couple meeting the description was arrested a few days after the crime, but no physical evidence was found. A mathematician calculated the probability that a randomly selected couple would possess the described characteristics as about 8.3×10^{-8}, or 1 in 12 million. Faced with such overwhelming odds and no physical evidence, the jury decided that the defendants must have been the only such couple and convicted them. The Supreme Court thought that a more useful probability should have been calculated. Based on the testimony of the witnesses, there was a couple that met the above description. Given that there was already one couple who met the description, what is the conditional probability that there was also a second couple such as the defendants?

Let p be the probability that a randomly selected couple from a population of n couples has certain characteristics. Let A be the event that at least one couple in the population has the characteristics, and let B be the event that at least two couples

have the characteristics. What we seek is $\Pr(B|A)$. Since $B \subset A$, it follows that

$$\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)} = \frac{\Pr(B)}{\Pr(A)}.$$

We shall calculate $\Pr(B)$ and $\Pr(A)$ by breaking each event into more manageable pieces. Suppose that we number the n couples in the population from 1 to n. Let A_i be the event that couple number i has the characteristics in question for $i = 1, \dots, n$, and let C be the event that exactly one couple has the characteristics. Then

$$A = (A_1^c \cap A_2^c \cdots \cap A_n^c)^c,$$

$$C = (A_1 \cap A_2^c \cdots \cap A_n^c) \cup (A_1^c \cap A_2 \cap A_3^c \cdots \cap A_n^c) \cup \cdots \cup (A_1^c \cap \cdots \cap A_{n-1}^c \cap A_n),$$

$$B = A \cap C^c.$$

Assuming that the n couples are mutually independent, $\Pr(A^c) = (1-p)^n$, and $\Pr(A) = 1 - (1-p)^n$. The n events whose union is C are disjoint and each one has probability $p(1-p)^{n-1}$, so $\Pr(C) = np(1-p)^{n-1}$. Since $A = B \cup C$ with B and C disjoint, we have

$$\Pr(B) = \Pr(A) - \Pr(C) = 1 - (1-p)^n - np(1-p)^{n-1}.$$

So,

$$\Pr(B|A) = \frac{1 - (1-p)^n - np(1-p)^{n-1}}{1 - (1-p)^n}. \tag{2.2.3}$$

The Supreme Court of California reasoned that, since the crime occurred in a heavily populated area, n would be in the millions. For example, with $p = 8.3 \times 10^{-8}$ and $n = 8{,}000{,}000$, the value of (2.2.3) is 0.2966. Such a probability suggests that there is a reasonable chance that there was another couple meeting the same description as the witnesses provided. Of course, the court did not know how large n was, but the fact that (2.2.3) could easily be so large was grounds enough to rule that reasonable doubt remained as to the guilt of the defendants. ◀

Independence and Conditional Probability Two events A and B with positive probability are independent if and only if $\Pr(A|B) = \Pr(A)$. Similar results hold for larger collections of independent events. The following theorem, for example, is straightforward to prove based on the definition of independence.

Theorem 2.2.2 Let $A_1, \dots, A_k$ be events such that $\Pr(A_1 \cap \cdots \cap A_k) > 0$. Then $A_1, \dots, A_k$ are independent if and only if, for every two disjoint subsets $\{i_1, \dots, i_m\}$ and $\{j_1, \dots, j_\ell\}$ of $\{1, \dots, k\}$, we have

$$\Pr(A_{i_1} \cap \cdots \cap A_{i_m} | A_{j_1} \cap \cdots \cap A_{j_\ell}) = \Pr(A_{i_1} \cap \cdots \cap A_{i_m}). \quad \blacksquare$$

Theorem 2.2.2 says that k events are independent if and only if learning that some of the events occur does not change the probability that any combination of the other events occurs.

The Meaning of Independence We have given a mathematical definition of independent events in Definition 2.2.1. We have also given some interpretations for what it means for events to be independent. The most instructive interpretation is the one based on conditional probability. If learning that B occurs does not change the probability of A, then A and B are independent. In simple examples such as tossing what we believe to be a fair coin, we would generally not expect to change our minds

about what is likely to happen on later flips after we observe earlier flips; hence, we declare the events that concern different flips to be independent. However, consider a situation similar to Example 2.2.5 in which items produced by a machine are inspected to see whether or not they are defective. In Example 2.2.5, we declared that the different items were independent and that each item had probability p of being defective. This might make sense if we were confident that we knew how well the machine was performing. But if we were unsure of how the machine were performing, we could easily imagine changing our mind about the probability that the 10th item is defective depending on how many of the first nine items are defective. To be specific, suppose that we begin by thinking that the probability is 0.08 that an item will be defective. If we observe one or zero defective items in the first nine, we might not make much revision to the probability that the 10th item is defective. On the other hand, if we observe eight or nine defectives in the first nine items, we might be uncomfortable keeping the probability at 0.08 that the 10th item will be defective. In summary, when deciding whether to model events as independent, try to answer the following question: "If I were to learn that some of these events occurred, would I change the probabilities of any of the others?" If we feel that we already know everything that we could learn from these events about how likely the others should be, we can safely model them as independent. If, on the other hand, we feel that learning some of these events could change our minds about how likely some of the others are, then we should be more careful about determining the conditional probabilities and not model the events as independent.

Mutually Exclusive Events and Mutually Independent Events Two similar-sounding definitions have appeared earlier in this text. Definition 1.4.10 defines mutually exclusive events, and Definition 2.2.2 defines mutually independent events. It is almost never the case that the same set of events satisfies both definitions. The reason is that if events are disjoint (mutually exclusive), then learning that one occurs means that the others definitely did not occur. Hence, learning that one occurs would change the probabilities for all the others to 0, unless the others already had probability 0. Indeed, this suggests the only condition in which the two definitions would both apply to the same collection of events. The proof of the following result is left to Exercise 24 in this section.

Theorem 2.2.3 Let $n > 1$ and let $A_1, \ldots, A_n$ be events that are mutually exclusive. The events are also mutually independent if and only if all the events except possibly one of them has probability 0. ■

Conditionally Independent Events

Conditional probability and independence combine into one of the most versatile models of data collection. The idea is that, in many circumstances, we are unwilling to say that certain events are independent because we believe that learning some of them will provide information about how likely the others are to occur. But if we knew the frequency with which such events would occur, we might then be willing to assume that they are independent. This model can be illustrated using one of the examples from earlier in this section.

Example 2.2.10 Inspecting Items. Consider again the situation in Example 2.2.5. This time, however, suppose that we believe that we would change our minds about the probabilities of later items being defective were we to learn that certain numbers of early items

were defective. Suppose that we think of the number p from Example 2.2.5 as the proportion of defective items that we would expect to see if we were to inspect a very large sample of items. If we knew this proportion p, and if we were to sample only a few, say, six or 10 items now, we might feel confident maintaining that the probability of a later item being defective remains p even after we inspect some of the earlier items. On the other hand, if we are not sure what would be the proportion of defective items in a large sample, we might not feel confident keeping the probability the same as we continue to inspect.

To be precise, suppose that we treat the proportion p of defective items as unknown and that we are dealing with an augmented experiment as described in Definition 2.1.3. For simplicity, suppose that p can take one of two values, either 0.01 or 0.4, the first corresponding to normal operation and the second corresponding to a need for maintenance. Let B_1 be the event that $p = 0.01$, and let B_2 be the event that $p = 0.4$. If we knew that B_1 had occurred, then we would proceed under the assumption that the events $D_1, D_2, \ldots$ were independent with $\Pr(D_i|B_1) = 0.01$ for all i. For example, we could do the same calculations as in Examples 2.2.5 and 2.2.8 with $p = 0.01$. Let A be the event that we observe exactly two defectives in a random sample of six items. Then $\Pr(A|B_1) = \binom{6}{2}0.01^2 0.99^4 = 1.44 \times 10^{-3}$. Similarly, if we knew that B_2 had occurred, then we would assume that $D_1, D_2, \ldots$ were independent with $\Pr(D_i|B_2) = 0.4$. In this case, $\Pr(A|B_2) = \binom{6}{2}0.4^2 0.6^4 = 0.311$. ◀

In Example 2.2.10, there is no reason that p must be required to assume at most two different values. We could easily allow p to take a third value or a fourth value, etc. Indeed, in Chapter 3 we shall learn how to handle the case in which every number between 0 and 1 is a possible value of p. The point of the simple example is to illustrate the concept of assuming that events are independent conditional on another event, such as B_1 or B_2 in the example.

The formal concept illustrated in Example 2.2.10 is the following:

Definition 2.2.3 Conditional Independence. We say that events $A_1, \ldots, A_k$ are *conditionally independent given* B if, for every subcollection $A_{i_1}, \ldots, A_{i_j}$ of j of these events ($j = 2, 3, \ldots, k$),

$$\Pr\left(A_{i_1} \cap \cdots \cap A_{i_j} \middle| B\right) = \Pr(A_{i_1}|B) \cdots \Pr(A_{i_j}|B).$$

Definition 2.2.3 is identical to Definition 2.2.2 for independent events with the modification that *all* probabilities in the definition are now conditional on B. As a note, even if we assume that events $A_1, \ldots, A_k$ are conditionally independent given B, it is *not* necessary that they be conditionally independent given B^c. In Example 2.2.10, the events $D_1, D_2, \ldots$ were conditionally independent given both B_1 and $B_2 = B_1^c$, which is the typical situation. Exercise 16 in Sec. 2.3 is an example in which events are conditionally independent given one event B but are not conditionally independent given the complement B^c.

Recall that two events A_1 and A_2 (with $\Pr(A_1) > 0$) are independent if and only if $\Pr(A_2|A_1) = \Pr(A_2)$. A similar result holds for conditionally independent events.

Theorem 2.2.4 Suppose that A_1, A_2, and B are events such that $\Pr(A_1 \cap B) > 0$. Then A_1 and A_2 are conditionally independent given B if and only if $\Pr(A_2|A_1 \cap B) = \Pr(A_2|B)$. ■

This is another example of the claim we made earlier that every result we can prove has an analog conditional on an event B. The reader can prove this theorem in Exercise 22.

The Collector's Problem

Suppose that n balls are thrown in a random manner into r boxes ($r \le n$). We shall assume that the n throws are independent and that each of the r boxes is equally likely to receive any given ball. The problem is to determine the probability p that every box will receive at least one ball. This problem can be reformulated in terms of a collector's problem as follows: Suppose that each package of bubble gum contains the picture of a baseball player, that the pictures of r different players are used, that the picture of each player is equally likely to be placed in any given package of gum, and that pictures are placed in different packages independently of each other. The problem now is to determine the probability p that a person who buys n packages of gum ($n \ge r$) will obtain a complete set of r different pictures.

For $i = 1, \dots, r$, let A_i denote the event that the picture of player i is missing from all n packages. Then $\bigcup_{i=1}^{r} A_i$ is the event that the picture of at least one player is missing. We shall find $\Pr(\bigcup_{i=1}^{r} A_i)$ by applying Eq. (1.10.6).

Since the picture of each of the r players is equally likely to be placed in any particular package, the probability that the picture of player i will not be obtained in any particular package is $(r-1)/r$. Since the packages are filled independently, the probability that the picture of player i will not be obtained in any of the n packages is $[(r-1)/r]^n$. Hence,

$$\Pr(A_i) = \left(\frac{r-1}{r}\right)^n \quad \text{for } i = 1, \dots, r.$$

Now consider any two players i and j. The probability that neither the picture of player i nor the picture of player j will be obtained in any particular package is $(r-2)/r$. Therefore, the probability that neither picture will be obtained in any of the n packages is $[(r-2)/r]^n$. Thus,

$$\Pr(A_i \cap A_j) = \left(\frac{r-2}{r}\right)^n .$$

If we next consider any three players i, j, and k, we find that

$$\Pr(A_i \cap A_j \cap A_k) = \left(\frac{r-3}{r}\right)^n .$$

By continuing in this way, we finally arrive at the probability $\Pr(A_1 \cap A_2 \cap \cdots \cap A_r)$ that the pictures of all r players are missing from the n packages. Of course, this probability is 0. Therefore, by Eq. (1.10.6) of Sec. 1.10,

$$\begin{aligned}\Pr\left(\bigcup_{i=1}^{r} A_i\right) &= r\left(\frac{r-1}{r}\right)^n - \binom{r}{2}\left(\frac{r-2}{r}\right)^n + \cdots + (-1)^r\binom{r}{r-1}\left(\frac{1}{r}\right)^n \\ &= \sum_{j=1}^{r-1}(-1)^{j+1}\binom{r}{j}\left(1-\frac{j}{r}\right)^n.\end{aligned}$$

Since the probability p of obtaining a complete set of r different pictures is equal to $1 - \Pr(\bigcup_{i=1}^{r} A_i)$, it follows from the foregoing derivation that p can be written in the form

$$p = \sum_{j=0}^{r-1}(-1)^j\binom{r}{j}\left(1-\frac{j}{r}\right)^n.$$

Summary

A collection of events is independent if and only if learning that some of them occur does not change the probabilities that any combination of the rest of them occurs. Equivalently, a collection of events is independent if and only if the probability of the intersection of every subcollection is the product of the individual probabilities. The concept of independence has a version conditional on another event. A collection of events is independent conditional on B if and only if the conditional probability of the intersection of every subcollection given B is the product of the individual conditional probabilities given B. Equivalently, a collection of events is conditionally independent given B if and only if learning that some of them (and B) occur does not change the conditional probabilities given B that any combination of the rest of them occur. The full power of conditional independence will become more apparent after we introduce Bayes' theorem in the next section.

Exercises

1. If A and B are independent events and $\Pr(B) < 1$, what is the value of $\Pr(A^c|B^c)$?

2. Assuming that A and B are independent events, prove that the events A^c and B^c are also independent.

3. Suppose that A is an event such that $\Pr(A) = 0$ and that B is any other event. Prove that A and B are independent events.

4. Suppose that a person rolls two balanced dice three times in succession. Determine the probability that on each of the three rolls, the sum of the two numbers that appear will be 7.

5. Suppose that the probability that the control system used in a spaceship will malfunction on a given flight is 0.001. Suppose further that a duplicate, but completely independent, control system is also installed in the spaceship to take control in case the first system malfunctions. Determine the probability that the spaceship will be under the control of either the original system or the duplicate system on a given flight.

6. Suppose that 10,000 tickets are sold in one lottery and 5000 tickets are sold in another lottery. If a person owns 100 tickets in each lottery, what is the probability that she will win at least one first prize?

7. Two students A and B are both registered for a certain course. Assume that student A attends class 80 percent of the time, student B attends class 60 percent of the time, and the absences of the two students are independent.

- **a.** What is the probability that at least one of the two students will be in class on a given day?
- **b.** If at least one of the two students is in class on a given day, what is the probability that A is in class that day?

8. If three balanced dice are rolled, what is the probability that all three numbers will be the same?

9. Consider an experiment in which a fair coin is tossed until a head is obtained for the first time. If this experiment is performed three times, what is the probability that exactly the same number of tosses will be required for each of the three performances?

10. The probability that any child in a certain family will have blue eyes is 1/4, and this feature is inherited independently by different children in the family. If there are five children in the family and it is known that at least one of these children has blue eyes, what is the probability that at least three of the children have blue eyes?

11. Consider the family with five children described in Exercise 10.

- **a.** If it is known that the youngest child in the family has blue eyes, what is the probability that at least three of the children have blue eyes?
- **b.** Explain why the answer in part (a) is different from the answer in Exercise 10.

12. Suppose that A, B, and C are three independent events such that $\Pr(A) = 1/4$, $\Pr(B) = 1/3$, and $\Pr(C) = 1/2$. **(a)** Determine the probability that none of these three events will occur. **(b)** Determine the probability that exactly one of these three events will occur.

13. Suppose that the probability that any particle emitted by a radioactive material will penetrate a certain shield is 0.01. If 10 particles are emitted, what is the probability that exactly one of the particles will penetrate the shield?

14. Consider again the conditions of Exercise 13. If 10 particles are emitted, what is the probability that at least one of the particles will penetrate the shield?

15. Consider again the conditions of Exercise 13. How many particles must be emitted in order for the probability to be at least 0.8 that at least one particle will penetrate the shield?

16. In the World Series of baseball, two teams A and B play a sequence of games against each other, and the first team that wins a total of four games becomes the winner of the World Series. If the probability that team A will win any particular game against team B is 1/3, what is the probability that team A will win the World Series?

17. Two boys A and B throw a ball at a target. Suppose that the probability that boy A will hit the target on any throw is 1/3 and the probability that boy B will hit the target on any throw is 1/4. Suppose also that boy A throws first and the two boys take turns throwing. Determine the probability that the target will be hit for the first time on the third throw of boy A.

18. For the conditions of Exercise 17, determine the probability that boy A will hit the target before boy B does.

19. A box contains 20 red balls, 30 white balls, and 50 blue balls. Suppose that 10 balls are selected at random one at a time, with replacement; that is, each selected ball is replaced in the box before the next selection is made. Determine the probability that at least one color will be missing from the 10 selected balls.

20. Suppose that $A_1, \dots, A_k$ form a sequence of k independent events. Let $B_1, \dots, B_k$ be another sequence of k events such that for each value of j $(j = 1, \dots, k)$, either $B_j = A_j$ or $B_j = A_j^c$. Prove that $B_1, \dots, B_k$ are also independent events. *Hint:* Use an induction argument based on the number of events B_j for which $B_j = A_j^c$.

21. Prove Theorem 2.2.2 on page 71. *Hint:* The "only if" direction is direct from the definition of independence on page 68. For the "if" direction, use induction on the value of j in the definition of independence. Let $m = j - 1$ and let $\ell = 1$ with $j_1 = i_j$.

22. Prove Theorem 2.2.4 on page 73.

23. A programmer is about to attempt to compile a series of 11 similar programs. Let A_i be the event that the ith program compiles successfully for $i = 1, \dots, 11$. When the programming task is easy, the programmer expects that 80 percent of programs should compile. When the programming task is difficult, she expects that only 40 percent of the programs will compile. Let B be the event that the programming task was easy. The programmer believes that the events $A_1, \dots, A_{11}$ are conditionally independent given B and given B^c.

a. Compute the probability that exactly 8 out of 11 programs will compile given B.

b. Compute the probability that exactly 8 out of 11 programs will compile given B^c.

24. Prove Theorem 2.2.3 on page 72.

2.3 Bayes' Theorem

Suppose that we are interested in which of several disjoint events $B_1, \dots, B_k$ will occur and that we will get to observe some other event A. If $\Pr(A|B_i)$ *is available for each i, then Bayes' theorem is a useful formula for computing the conditional probabilities of the B_i events given A.*

We begin with a typical example.

Example 2.3.1 Test for a Disease. Suppose that you are walking down the street and notice that the Department of Public Health is giving a free medical test for a certain disease. The test is 90 percent reliable in the following sense: If a person has the disease, there is a probability of 0.9 that the test will give a positive response; whereas, if a person does not have the disease, there is a probability of only 0.1 that the test will give a positive response.

Data indicate that your chances of having the disease are only 1 in 10,000. However, since the test costs you nothing, and is fast and harmless, you decide to stop and take the test. A few days later you learn that you had a positive response to the test. Now, what is the probability that you have the disease? ◀

The last question in Example 2.3.1 is a prototype of the question for which Bayes' theorem was designed. We have at least two disjoint events ("you have the disease" and "you do not have the disease") about which we are uncertain, and we learn a piece of information (the result of the test) that tells us something about the uncertain events. Then we need to know how to revise the probabilities of the events in the light of the information we learned.

We now present the general structure in which Bayes' theorem operates before returning to the example.

Statement, Proof, and Examples of Bayes' Theorem

Example 2.3.2

Selecting Bolts. Consider again the situation in Example 2.1.8, in which a bolt is selected at random from one of two boxes. Suppose that we cannot tell without making a further effort from which of the two boxes the one bolt is being selected. For example, the boxes may be identical in appearance or somebody else may actually select the box, but we only get to see the bolt. Prior to selecting the bolt, it was equally likely that each of the two boxes would be selected. However, if we learn that event A has occurred, that is, a long bolt was selected, we can compute the conditional probabilities of the two boxes given A. To remind the reader, B_1 is the event that the box is selected containing 60 long bolts and 40 short bolts, while B_2 is the event that the box is selected containing 10 long bolts and 20 short bolts. In Example 2.1.9, we computed $\Pr(A) = 7/15$, $\Pr(A|B_1) = 3/5$, $\Pr(A|B_2) = 1/3$, and $\Pr(B_1) = \Pr(B_2) = 1/2$. So, for example,

$$\Pr(B_1|A) = \frac{\Pr(A \cap B_1)}{\Pr(A)} = \frac{\Pr(B_1)\Pr(A|B_1)}{\Pr(A)} = \frac{\frac{1}{2} \times \frac{3}{5}}{\frac{7}{15}} = \frac{9}{14}.$$

Since the first box has a higher proportion of long bolts than the second box, it seems reasonable that the probability of B_1 should rise after we learn that a long bolt was selected. It must be that $\Pr(B_2|A) = 5/14$ since one or the other box had to be selected. ◀

In Example 2.3.2, we started with uncertainty about which of two boxes would be chosen and then we observed a long bolt drawn from the chosen box. Because the two boxes have different chances of having a long bolt drawn, the observation of a long bolt changed the probabilities of each of the two boxes having been chosen. The precise calculation of how the probabilities change is the purpose of Bayes' theorem.

Theorem 2.3.1

Bayes' theorem. Let the events $B_1, \dots, B_k$ form a partition of the space S such that $\Pr(B_j) > 0$ for $j = 1, \dots, k$, and let A be an event such that $\Pr(A) > 0$. Then, for $i = 1, \dots, k$,

$$\Pr(B_i|A) = \frac{\Pr(B_i)\Pr(A|B_i)}{\sum_{j=1}^{k}\Pr(B_j)\Pr(A|B_j)}. \tag{2.3.1}$$

Proof By the definition of conditional probability,

$$\Pr(B_i|A) = \frac{\Pr(B_i \cap A)}{\Pr(A)}.$$

The numerator on the right side of Eq. (2.3.1) is equal to $\Pr(B_i \cap A)$ by Theorem 2.1.1. The denominator is equal to $\Pr(A)$ according to Theorem 2.1.4. ■

Example 2.3.3

Test for a Disease. Let us return to the example with which we began this section. We have just received word that we have tested positive for a disease. The test was 90 percent reliable in the sense that we described in Example 2.3.1. We want to know the probability that we have the disease after we learn that the result of the test is positive. Some readers may feel that this probability should be about 0.9. However, this feeling completely ignores the small probability of 0.0001 that you had the disease before taking the test. We shall let B_1 denote the event that you have the disease, and let B_2 denote the event that you do not have the disease. The events B_1 and B_2 form a partition. Also, let A denote the event that the response to the test is positive. The event A is information we will learn that tells us something about the partition elements. Then, by Bayes' theorem,

$$\Pr(B_1|A) = \frac{\Pr(A|B_1)\Pr(B_1)}{\Pr(A|B_1)\Pr(B_1) + \Pr(A|B_2)\Pr(B_2)}$$
$$= \frac{(0.9)(0.0001)}{(0.9)(0.0001) + (0.1)(0.9999)} = 0.00090.$$

Thus, the conditional probability that you have the disease given the test result is approximately only 1 in 1000. Of course, this conditional probability is approximately 9 times as great as the probability was before you were tested, but even the conditional probability is quite small.

Another way to explain this result is as follows: Only one person in every 10,000 actually has the disease, but the test gives a positive response for approximately one person in every 10. Hence, the number of positive responses is approximately 1000 times the number of persons who actually have the disease. In other words, out of every 1000 persons for whom the test gives a positive response, only one person actually has the disease. This example illustrates not only the use of Bayes' theorem but also the importance of taking into account all of the information available in a problem. ◀

Example 2.3.4

Identifying the Source of a Defective Item. Three different machines M_1, M_2, and M_3 were used for producing a large batch of similar manufactured items. Suppose that 20 percent of the items were produced by machine M_1, 30 percent by machine M_2, and 50 percent by machine M_3. Suppose further that 1 percent of the items produced by machine M_1 are defective, that 2 percent of the items produced by machine M_2 are defective, and that 3 percent of the items produced by machine M_3 are defective. Finally, suppose that one item is selected at random from the entire batch and it is found to be defective. We shall determine the probability that this item was produced by machine M_2.

Let B_i be the event that the selected item was produced by machine M_i ($i = 1, 2, 3$), and let A be the event that the selected item is defective. We must evaluate the conditional probability $\Pr(B_2|A)$.

The probability $\Pr(B_i)$ that an item selected at random from the entire batch was produced by machine M_i is as follows, for $i = 1, 2, 3$:

$$\Pr(B_1) = 0.2, \quad \Pr(B_2) = 0.3, \quad \Pr(B_3) = 0.5.$$

Furthermore, the probability $\Pr(A|B_i)$ that an item produced by machine M_i will be defective is

$$\Pr(A|B_1) = 0.01, \quad \Pr(A|B_2) = 0.02, \quad \Pr(A|B_3) = 0.03.$$

It now follows from Bayes' theorem that

$$\Pr(B_2|A) = \frac{\Pr(B_2)\Pr(A|B_2)}{\sum_{j=1}^{3}\Pr(B_j)\Pr(A|B_j)}$$
$$= \frac{(0.3)(0.02)}{(0.2)(0.01)+(0.3)(0.02)+(0.5)(0.03)} = 0.26. \blacktriangleleft$$

Example 2.3.5

Identifying Genotypes. Consider a gene that has two alleles (see Example 1.6.4 on page 23) A and a. Suppose that the gene exhibits itself through a trait (such as hair color or blood type) with two versions. We call A *dominant* and a *recessive* if individuals with genotypes AA and Aa have the same version of the trait and the individuals with genotype aa have the other version. The two versions of the trait are called *phenotypes*. We shall call the phenotype exhibited by individuals with genotypes AA and Aa the *dominant trait,* and the other trait will be called the *recessive trait*. In population genetics studies, it is common to have information on the phenotypes of individuals, but it is rather difficult to determine genotypes. However, some information about genotypes can be obtained by observing phenotypes of parents and children.

Assume that the allele A is dominant, that individuals mate independently of genotype, and that the genotypes AA, Aa, and aa occur in the population with probabilities 1/4, 1/2, and 1/4, respectively. We are going to observe an individual whose parents are not available, and we shall observe the phenotype of this individual. Let E be the event that the observed individual has the dominant trait. We would like to revise our opinion of the possible genotypes of the parents. There are six possible genotype combinations, $B_1, \ldots, B_6$, for the parents prior to making any observations, and these are listed in Table 2.2.

The probabilities of the B_i were computed using the assumption that the parents mated independently of genotype. For example, B_3 occurs if the father is AA and the mother is aa (probability 1/16) or if the father is aa and the mother is AA (probability 1/16). The values of $\Pr(E|B_i)$ were computed assuming that the two available alleles are passed from parents to children with probability 1/2 each and independently for the two parents. For example, given B_4, the event E occurs if and only if the child does not get two a's. The probability of getting a from both parents given B_4 is 1/4, so $\Pr(E|B_4) = 3/4$.

Now we shall compute $\Pr(B_1|E)$ and $\Pr(B_5|E)$. We leave the other calculations to the reader. The denominator of Bayes' theorem is the same for both calculations, namely,

$$\Pr(E) = \sum_{i=1}^{5}\Pr(B_i)\Pr(E|B_i)$$
$$= \frac{1}{16}\times 1 + \frac{1}{4}\times 1 + \frac{1}{8}\times 1 + \frac{1}{4}\times\frac{3}{4} + \frac{1}{4}\times\frac{1}{2} + \frac{1}{16}\times 0 = \frac{3}{4}.$$

Table 2.2 Parental genotypes for Example 2.3.5

	(AA, AA)	(AA, Aa)	(AA, aa)	(Aa, Aa)	(Aa, aa)	(aa, aa)
Name of event	B_1	B_2	B_3	B_4	B_5	B_6
Probability of B_i	1/16	1/4	1/8	1/4	1/4	1/16
$\Pr(E\|B_i)$	1	1	1	3/4	1/2	0

Applying Bayes' theorem, we get

$$\Pr(B_1|E) = \frac{\frac{1}{16} \times 1}{\frac{3}{4}} = \frac{1}{12}, \quad \Pr(B_5|E) = \frac{\frac{1}{4} \times \frac{1}{2}}{\frac{3}{4}} = \frac{1}{6}. \blacktriangleleft$$

Note: Conditional Version of Bayes' Theorem. There is also a version of Bayes' theorem conditional on an event C:

$$\Pr(B_i|A \cap C) = \frac{\Pr(B_i|C)\Pr(A|B_i \cap C)}{\sum_{j=1}^{k} \Pr(B_j|C)\Pr(A|B_j \cap C)}. \tag{2.3.2}$$

Prior and Posterior Probabilities

In Example 2.3.4, a probability like $\Pr(B_2)$ is often called the *prior probability* that the selected item will have been produced by machine M_2, because $\Pr(B_2)$ is the probability of this event before the item is selected and before it is known whether the selected item is defective or nondefective. A probability like $\Pr(B_2|A)$ is then called the *posterior probability* that the selected item was produced by machine M_2, because it is the probability of this event after it is known that the selected item is defective.

Thus, in Example 2.3.4, the prior probability that the selected item will have been produced by machine M_2 is 0.3. After an item has been selected and has been found to be defective, the posterior probability that the item was produced by machine M_2 is 0.26. Since this posterior probability is smaller than the prior probability that the item was produced by machine M_2, the posterior probability that the item was produced by one of the other machines must be larger than the prior probability that it was produced by one of those machines (see Exercises 1 and 2 at the end of this section).

Computation of Posterior Probabilities in More Than One Stage

Suppose that a box contains one fair coin and one coin with a head on each side. Suppose also that one coin is selected at random and that when it is tossed, a head is obtained. We shall determine the probability that the coin is the fair coin.

Let B_1 be the event that the coin is fair, let B_2 be the event that the coin has two heads, and let H_1 be the event that a head is obtained when the coin is tossed. Then, by Bayes' theorem,

$$\begin{aligned}\Pr(B_1|H_1) &= \frac{\Pr(B_1)\Pr(H_1|B_1)}{\Pr(B_1)\Pr(H_1|B_1) + \Pr(B_2)\Pr(H_1|B_2)} \\ &= \frac{(1/2)(1/2)}{(1/2)(1/2) + (1/2)(1)} = \frac{1}{3}.\end{aligned} \tag{2.3.3}$$

Thus, after the first toss, the posterior probability that the coin is fair is 1/3.

Now suppose that the same coin is tossed again and we assume that the two tosses are conditionally independent given both B_1 and B_2. Suppose that another head is obtained. There are two ways of determining the new value of the posterior probability that the coin is fair.

The first way is to return to the beginning of the experiment and assume again that the prior probabilities are $\Pr(B_1) = \Pr(B_2) = 1/2$. We shall let $H_1 \cap H_2$ denote the event in which heads are obtained on two tosses of the coin, and we shall calculate the posterior probability $\Pr(B_1|H_1 \cap H_2)$ that the coin is fair after we have observed the

event $H_1 \cap H_2$. The assumption that the tosses are conditionally independent given B_1 means that $\Pr(H_1 \cap H_2|B_1) = 1/2 \times 1/2 = 1/4$. By Bayes' theorem,

$$\Pr(B_1|H_1 \cap H_2) = \frac{\Pr(B_1)\Pr(H_1 \cap H_2|B_1)}{\Pr(B_1)\Pr(H_1 \cap H_2|B_1) + \Pr(B_2)\Pr(H_1 \cap H_2|B_2)}$$
$$= \frac{(1/2)(1/4)}{(1/2)(1/4) + (1/2)(1)} = \frac{1}{5}. \tag{2.3.4}$$

The second way of determining this same posterior probability is to use the conditional version of Bayes' theorem (2.3.2) given the event H_1. Given H_1, the conditional probability of B_1 is 1/3, and the conditional probability of B_2 is therefore 2/3. These conditional probabilities can now serve as the prior probabilities for the next stage of the experiment, in which the coin is tossed a second time. Thus, we can apply (2.3.2) with $C = H_1$, $\Pr(B_1|H_1) = 1/3$, and $\Pr(B_2|H_1) = 2/3$. We can then compute the posterior probability $\Pr(B_1|H_1 \cap H_2)$ that the coin is fair after we have observed a head on the second toss and a head on the first toss. We shall need $\Pr(H_2|B_1 \cap H_1)$, which equals $\Pr(H_2|B_1) = 1/2$ by Theorem 2.2.4 since H_1 and H_2 are conditionally independent given B_1. Since the coin is two-headed when B_2 occurs, $\Pr(H_2|B_2 \cap H_1) = 1$. So we obtain

$$\Pr(B_1|H_1 \cap H_2) = \frac{\Pr(B_1|H_1)\Pr(H_2|B_1 \cap H_1)}{\Pr(B_1|H_1)\Pr(H_2|B_1 \cap H_1) + \Pr(B_2|H_1)\Pr(H_2|B_2 \cap H_1)}$$
$$= \frac{(1/3)(1/2)}{(1/3)(1/2) + (2/3)(1)} = \frac{1}{5}. \tag{2.3.5}$$

The posterior probability of the event B_1 obtained in the second way is the same as that obtained in the first way. We can make the following general statement: If an experiment is carried out in more than one stage, then the posterior probability of every event can also be calculated in more than one stage. After each stage has been carried out, the posterior probability calculated for the event after that stage serves as the prior probability for the next stage. The reader should look back at (2.3.2) to see that this interpretation is precisely what the conditional version of Bayes' theorem says. The example we have been doing with coin tossing is typical of many applications of Bayes' theorem and its conditional version because we are assuming that the observable events are conditionally independent given each element of the partition $B_1, \ldots, B_k$ (in this case, $k = 2$). The conditional independence makes the probability of H_i (head on ith toss) given B_1 (or given B_2) the same whether or not we also condition on earlier tosses (see Theorem 2.2.4).

Conditionally Independent Events

The calculations that led to (2.3.3) and (2.3.5) together with Example 2.2.10 illustrate simple cases of a very powerful statistical model for observable events. It is very common to encounter a sequence of events that we believe are similar in that they all have the same probability of occurring. It is also common that the order in which the events are labeled does not affect the probabilities that we assign. However, we often believe that these events are not independent, because, if we were to observe some of them, we would change our minds about the probability of the ones we had not observed depending on how many of the observed events occur. For example, in the coin-tossing calculation leading up to Eq. (2.3.3), before any tosses occur, the probability of H_2 is the same as the probability of H_1, namely, the

denominator of (2.3.3), 3/4, as Theorem 2.1.4 says. However, after observing that the event H_1 occurs, the probability of H_2 is $\Pr(H_2|H_1)$, which is the denominator of (2.3.5), 5/6, as computed by the conditional version of the law of total probability (2.1.5). Even though we might treat the coin tosses as independent conditional on the coin being fair, and we might treat them as independent conditional on the coin being two-headed (in which case we know what will happen every time anyway), we cannot treat them as independent without the conditioning information. The conditioning information removes an important source of uncertainty from the problem, so we partition the sample space accordingly. Now we can use the conditional independence of the tosses to calculate joint probabilities of various combinations of events conditionally on the partition events. Finally, we can combine these probabilities using Theorem 2.1.4 and (2.1.5). Two more examples will help to illustrate these ideas.

Example 2.3.6

Learning about a Proportion. In Example 2.2.10 on page 72, a machine produced defective parts in one of two proportions, $p = 0.01$ or $p = 0.4$. Suppose that the prior probability that $p = 0.01$ is 0.9. After sampling six parts at random, suppose that we observe two defectives. What is the posterior probability that $p = 0.01$?

Let $B_1 = \{p = 0.01\}$ and $B_2 = \{p = 0.4\}$ as in Example 2.2.10. Let A be the event that two defectives occur in a random sample of size six. The prior probability of B_1 is 0.9, and the prior probability of B_2 is 0.1. We already computed $\Pr(A|B_1) = 1.44 \times 10^{-3}$ and $\Pr(A|B_2) = 0.311$ in Example 2.2.10. Bayes' theorem tells us that

$$\Pr(B_1|A) = \frac{0.9 \times 1.44 \times 10^{-3}}{0.9 \times 1.44 \times 10^{-3} + 0.1 \times 0.311} = 0.04.$$

Even though we thought originally that B_1 had probability as high as 0.9, after we learned that there were two defective items in a sample as small as six, we changed our minds dramatically and now we believe that B_1 has probability as small as 0.04. The reason for this major change is that the event A that occurred has much higher probability if B_2 is true than if B_1 is true. ◄

Example 2.3.7

A Clinical Trial. Consider the same clinical trial described in Examples 2.1.12 and 2.1.13. Let E_i be the event that the ith patient has success as her outcome. Recall that B_j is the event that $p = (j - 1)/10$ for $j = 1, \ldots, 11$, where p is the proportion of successes among all possible patients. If we knew which B_j occurred, we would say that $E_1, E_2, \ldots$ were independent. That is, we are willing to model the patients as conditionally independent given each event B_j, and we set $\Pr(E_i|B_j) = (j - 1)/10$ for all i, j. We shall still assume that $\Pr(B_j) = 1/11$ for all j prior to the start of the trial. We are now in position to express what we learn about p by computing posterior probabilities for the B_j events after each patient finishes the trial.

For example, consider the first patient. We calculated $\Pr(E_1) = 1/2$ in (2.1.6). If E_1 occurs, we apply Bayes' theorem to get

$$\Pr(B_j|E_1) = \frac{\Pr(E_1|B_j)\Pr(B_j)}{1/2} = \frac{2(j-1)}{10 \times 11} = \frac{j-1}{55}. \tag{2.3.6}$$

After observing one success, the posterior probabilities of large values of p are higher than their prior probabilities and the posterior probabilities of low values of p are lower than their prior probabilities as we would expect. For example, $\Pr(B_1|E_1) = 0$, because $p = 0$ is ruled out after one success. Also, $\Pr(B_2|E_1) = 0.0182$, which is much smaller than its prior value 0.0909, and $\Pr(B_{11}|E_1) = 0.1818$, which is larger than its prior value 0.0909.

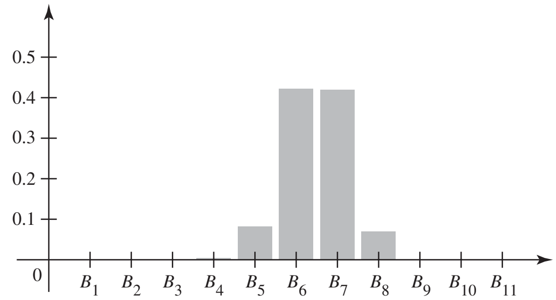


Figure 2.3 The posterior probabilities of partition elements after 40 patients in Example 2.3.7.

We could check how the posterior probabilities behave after each patient is observed. However, we shall skip ahead to the point at which all 40 patients in the imipramine column of Table 2.1 have been observed. Let A stand for the observed event that 22 of them are successes and 18 are failures. We can use the same reasoning as in Example 2.2.5 to compute $\Pr(A|B_j)$. There are $\binom{40}{22}$ possible sequences of 40 patients with 22 successes, and, conditional on B_j, the probability of each sequence is $([j-1]/10)^{22}(1-[j-1]/10)^{18}$.

So,

$$\Pr(A|B_j) = \binom{40}{22}([j-1]/10)^{22}(1-[j-1]/10)^{18}, \tag{2.3.7}$$

for each j. Then Bayes' theorem tells us that

$$\Pr(B_j|A) = \frac{\frac{1}{11}\binom{40}{22}([j-1]/10)^{22}(1-[j-1]/10)^{18}}{\sum_{i=1}^{11}\frac{1}{11}\binom{40}{22}([i-1]/10)^{22}(1-[i-1]/10)^{18}}.$$

Figure 2.3 shows the posterior probabilities of the 11 partition elements after observing A. Notice that the probabilities of B_6 and B_7 are the highest, 0.42. This corresponds to the fact that the proportion of successes in the observed sample is $22/40 = 0.55$, halfway between $(6-1)/10$ and $(7-1)/10$.

We can also compute the probability that the next patient will be a success both before the trial and after the 40 patients. Before the trial, $\Pr(E_{41}) = \Pr(E_1)$, which equals 1/2, as computed in (2.1.6). After observing the 40 patients, we can compute $\Pr(E_{41}|A)$ using the conditional version of the law of total probability, (2.1.5):

$$\Pr(E_{41}|A) = \sum_{j=1}^{11} \Pr(E_{41}|B_j \cap A)\Pr(B_j|A). \tag{2.3.8}$$

Using the values of $\Pr(B_j|A)$ in Fig. 2.3 and the fact that $\Pr(E_{41}|B_j \cap A) = \Pr(E_{41}|B_j) = (j-1)/10$ (conditional independence of the E_i given the B_j), we compute (2.3.8) to be 0.5476. This is also very close to the observed frequency of success. ◀

The calculation at the end of Example 2.3.7 is typical of what happens after observing many conditionally independent events with the same conditional probability of occurrence. The conditional probability of the next event given those that were observed tends to be close to the observed frequency of occurrence among the observed events. Indeed, when there is substantial data, the choice of prior probabilities becomes far less important.

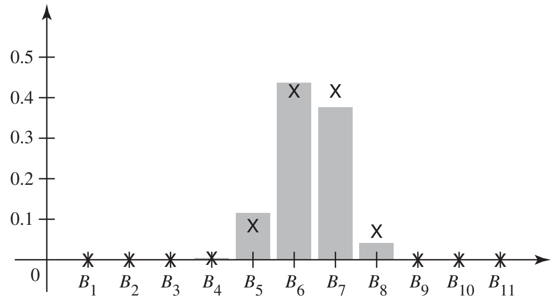


Figure 2.4 The posterior probabilities of partition elements after 40 patients in Example 2.3.8. The X characters mark the values of the posterior probabilities calculated in Example 2.3.7.

Example 2.3.8 The Effect of Prior Probabilities. Consider the same clinical trial as in Example 2.3.7. This time, suppose that a different researcher has a different prior opinion about the value of p, the probability of success. This researcher believes the following prior probabilities:

Event	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}
p	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Prior prob.	0.00	0.19	0.19	0.17	0.14	0.11	0.09	0.06	0.04	0.01	0.00

We can recalculate the posterior probabilities using Bayes' theorem, and we get the values pictured in Fig. 2.4. To aid comparison, the posterior probabilities from Example 2.3.7 are also plotted in Fig. 2.4 using the symbol X. One can see how close the two sets of posterior probabilities are despite the large differences between the prior probabilities. If there had been fewer patients observed, there would have been larger differences between the two sets of posterior probabilites because the observed events would have provided less information. (See Exercise 12 in this section.) ◂

Summary

Bayes' theorem tells us how to compute the conditional probability of each event in a partition given an observed event A. A major use of partitions is to divide the sample space into small enough pieces so that a collection of events of interest become conditionally independent given each event in the partition.

Exercises

1. Suppose that k events $B_1, \dots, B_k$ form a partition of the sample space S. For $i = 1, \dots, k$, let $\Pr(B_i)$ denote the prior probability of B_i. Also, for each event A such that $\Pr(A) > 0$, let $\Pr(B_i|A)$ denote the posterior probability of B_i given that the event A has occurred. Prove that if $\Pr(B_1|A) < \Pr(B_1)$, then $\Pr(B_i|A) > \Pr(B_i)$ for at least one value of i $(i = 2, \dots, k)$.

2. Consider again the conditions of Example 2.3.4 in this section, in which an item was selected at random from a batch of manufactured items and was found to be defective. For which values of i ($i = 1, 2, 3$) is the posterior probability that the item was produced by machine M_i larger than the prior probability that the item was produced by machine M_i?

3. Suppose that in Example 2.3.4 in this section, the item selected at random from the entire lot is found to be non-defective. Determine the posterior probability that it was produced by machine M_2.

4. A new test has been devised for detecting a particular type of cancer. If the test is applied to a person who has this type of cancer, the probability that the person will have a positive reaction is 0.95 and the probability that the person will have a negative reaction is 0.05. If the test is applied to a person who does not have this type of cancer, the probability that the person will have a positive reaction is 0.05 and the probability that the person will have a negative reaction is 0.95. Suppose that in the general population, one person out of every 100,000 people has this type of cancer. If a person selected at random has a positive reaction to the test, what is the probability that he has this type of cancer?

5. In a certain city, 30 percent of the people are Conservatives, 50 percent are Liberals, and 20 percent are Independents. Records show that in a particular election, 65 percent of the Conservatives voted, 82 percent of the Liberals voted, and 50 percent of the Independents voted. If a person in the city is selected at random and it is learned that she did not vote in the last election, what is the probability that she is a Liberal?

6. Suppose that when a machine is adjusted properly, 50 percent of the items produced by it are of high quality and the other 50 percent are of medium quality. Suppose, however, that the machine is improperly adjusted during 10 percent of the time and that, under these conditions, 25 percent of the items produced by it are of high quality and 75 percent are of medium quality.

- **a.** Suppose that five items produced by the machine at a certain time are selected at random and inspected. If four of these items are of high quality and one item is of medium quality, what is the probability that the machine was adjusted properly at that time?
- **b.** Suppose that one additional item, which was produced by the machine at the same time as the other five items, is selected and found to be of medium quality. What is the new posterior probability that the machine was adjusted properly?

7. Suppose that a box contains five coins and that for each coin there is a different probability that a head will be obtained when the coin is tossed. Let p_i denote the probability of a head when the ith coin is tossed ($i = 1, \ldots, 5$), and suppose that $p_1 = 0$, $p_2 = 1/4$, $p_3 = 1/2$, $p_4 = 3/4$, and $p_5 = 1$.

- **a.** Suppose that one coin is selected at random from the box and when it is tossed once, a head is obtained. What is the posterior probability that the ith coin was selected ($i = 1, \ldots, 5$)?
- **b.** If the same coin were tossed again, what would be the probability of obtaining another head?
- **c.** If a tail had been obtained on the first toss of the selected coin and the same coin were tossed again, what would be the probability of obtaining a head on the second toss?

8. Consider again the box containing the five different coins described in Exercise 7. Suppose that one coin is selected at random from the box and is tossed repeatedly until a head is obtained.

- **a.** If the first head is obtained on the fourth toss, what is the posterior probability that the ith coin was selected ($i = 1, \ldots, 5$)?
- **b.** If we continue to toss the same coin until another head is obtained, what is the probability that exactly three additional tosses will be required?

9. Consider again the conditions of Exercise 14 in Sec. 2.1. Suppose that several parts will be observed and that the different parts are conditionally independent given each of the three states of repair of the machine. If seven parts are observed and exactly one is defective, compute the posterior probabilities of the three states of repair.

10. Consider again the conditions of Example 2.3.5, in which the phenotype of an individual was observed and found to be the dominant trait. For which values of i ($i = 1, \ldots, 6$) is the posterior probability that the parents have the genotypes of event B_i smaller than the prior probability that the parents have the genotyes of event B_i?

11. Suppose that in Example 2.3.5 the observed individual has the recessive trait. Determine the posterior probability that the parents have the genotypes of event B_4.

12. In the clinical trial in Examples 2.3.7 and 2.3.8, suppose that we have only observed the first five patients and three of the five had been successes. Use the two different sets of prior probabilities from Examples 2.3.7 and 2.3.8 to calculate two sets of posterior probabilities. Are these two sets of posterior probabilities as close to each other as were the two in Examples 2.3.7 and 2.3.8? Why or why not?

13. Suppose that a box contains one fair coin and one coin with a head on each side. Suppose that a coin is drawn at random from this box and that we begin to flip the coin. In Eqs. (2.3.4) and (2.3.5), we computed the conditional

probability that the coin was fair given that the first two flips both produce heads.

a. Suppose that the coin is flipped a third time and another head is obtained. Compute the probability that the coin is fair given that all three flips produced heads.

b. Suppose that the coin is flipped a fourth time and the result is tails. Compute the posterior probability that the coin is fair.

14. Consider again the conditions of Exercise 23 in Sec. 2.2. Assume that $\Pr(B) = 0.4$. Let A be the event that exactly 8 out of 11 programs compiled. Compute the conditional probability of B given A.

15. Use the prior probabilities in Example 2.3.8 for the events $B_1, \dots, B_{11}$. Let E_1 be the event that the first patient is a success. Compute the probability of E_1 and explain why it is so much less than the value computed in Example 2.3.7.

16. Consider a machine that produces items in sequence. Under normal operating conditions, the items are independent with probability 0.01 of being defective. However, it is possible for the machine to develop a "memory" in the following sense: After each defective item, and independent of anything that happened earlier, the probability that the next item is defective is 2/5. After each nondefective item, and independent of anything that happened earlier, the probability that the next item is defective is 1/165.

Assume that the machine is either operating normally for the whole time we observe or has a memory for the whole time that we observe. Let B be the event that the machine is operating normally, and assume that $\Pr(B) = 2/3$. Let D_i be the event that the ith item inspected is defective. Assume that D_1 is independent of B.

a. Prove that $\Pr(D_i) = 0.01$ for all i. *Hint:* Use induction.

b. Assume that we observe the first six items and the event that occurs is $E = D_1^c \cap D_2^c \cap D_3 \cap D_4 \cap D_5^c \cap D_6^c$. That is, the third and fourth items are defective, but the other four are not. Compute $\Pr(B|D)$.

⋆ 2.4 The Gambler's Ruin Problem

Consider two gamblers with finite resources who repeatedly play the same game against each other. Using the tools of conditional probability, we can calculate the probability that each of the gamblers will eventually lose all of his money to the opponent.

Statement of the Problem

Suppose that two gamblers A and B are playing a game against each other. Let p be a given number $(0 < p < 1)$, and suppose that on each play of the game, the probability that gambler A will win one dollar from gambler B is p and the probability that gambler B will win one dollar from gambler A is $1 - p$. Suppose also that the initial fortune of gambler A is i dollars and the initial fortune of gambler B is $k - i$ dollars, where i and $k - i$ are given positive integers. Thus, the total fortune of the two gamblers is k dollars. Finally, suppose that the gamblers play the game repeatedly and independently until the fortune of one of them has been reduced to 0 dollars. Another way to think about this problem is that B is a casino and A is a gambler who is determined to quit as soon he wins $k - i$ dollars from the casino or when he goes broke, whichever comes first.

We shall now consider this game from the point of view of gambler A. His initial fortune is i dollars and on each play of the game his fortune will either increase by one dollar with a probability of p or decrease by one dollar with a probability of $1 - p$. If $p > 1/2$, the game is favorable to him; if $p < 1/2$, the game is unfavorable to him; and if $p = 1/2$, the game is equally favorable to both gamblers. The game ends either when the fortune of gambler A reaches k dollars, in which case gambler B will have no money left, or when the fortune of gambler A reaches 0 dollars. The problem is to

determine the probability that the fortune of gambler A will reach k dollars before it reaches 0 dollars. Because one of the gamblers will have no money left at the end of the game, this problem is called the *Gambler's Ruin* problem.

Solution of the Problem

We shall continue to assume that the total fortune of the gamblers A and B is k dollars, and we shall let a_i denote the probability that the fortune of gambler A will reach k dollars before it reaches 0 dollars, given that his initial fortune is i dollars. We assume that the game is the same each time it is played and the plays are independent of each other. It follows that, after each play, the Gambler's Ruin problem essentially starts over with the only change being that the initial fortunes of the two gamblers have changed. In particular, for each $j = 0, \dots, k$, each time that we observe a sequence of plays that lead to gambler A's fortune being j dollars, the conditional probability, given such a sequence, that gambler A wins is a_j. If gambler A's fortune ever reaches 0, then gambler A is ruined, hence $a_0 = 0$. Similarly, if his fortune ever reaches k, then gambler A has won, hence $a_k = 1$. We shall now determine the value of a_i for $i = 1, \dots, k - 1$.

Let A_1 denote the event that gambler A wins one dollar on the first play of the game, let B_1 denote the event that gambler A loses one dollar on the first play of the game, and let W denote the event that the fortune of gambler A ultimately reaches k dollars before it reaches 0 dollars. Then

$$\begin{aligned}\Pr(W) &= \Pr(A_1)\Pr(W|A_1) + \Pr(B_1)\Pr(W|B_1)\\ &= p\Pr(W|A_1) + (1-p)\Pr(W|B_1).\end{aligned} \tag{2.4.1}$$

Since the initial fortune of gambler A is i dollars ($i = 1, \dots, k - 1$), then $\Pr(W) = a_i$. Furthermore, if gambler A wins one dollar on the first play of the game, then his fortune becomes $i + 1$ dollars and the conditional probability $\Pr(W|A_1)$ that his fortune will ultimately reach k dollars is therefore a_{i+1}. If A loses one dollar on the first play of the game, then his fortune becomes $i - 1$ dollars and the conditional probability $\Pr(W|B_1)$ that his fortune will ultimately reach k dollars is therefore a_{i-1}. Hence, by Eq. (2.4.1),

$$a_i = pa_{i+1} + (1-p)a_{i-1}. \tag{2.4.2}$$

We shall let $i = 1, \dots, k - 1$ in Eq. (2.4.2). Then, since $a_0 = 0$ and $a_k = 1$, we obtain the following $k - 1$ equations:

$$\begin{aligned}a_1 &= pa_2,\\ a_2 &= pa_3 + (1-p)a_1,\\ a_3 &= pa_4 + (1-p)a_2,\\ &\vdots\\ a_{k-2} &= pa_{k-1} + (1-p)a_{k-3},\\ a_{k-1} &= p + (1-p)a_{k-2}.\end{aligned} \tag{2.4.3}$$

If the value of a_i on the left side of the ith equation is rewritten in the form $pa_i + (1-p)a_i$ and some elementary algebra is performed, then these $k - 1$ equations can

be rewritten as follows:

$$
\begin{aligned}
a_2 - a_1 &= \frac{1-p}{p} a_1, \\
a_3 - a_2 &= \frac{1-p}{p}(a_2 - a_1) &&= \left(\frac{1-p}{p}\right)^2 a_1, \\
a_4 - a_3 &= \frac{1-p}{p}(a_3 - a_2) &&= \left(\frac{1-p}{p}\right)^3 a_1, \\
&\vdots \\
a_{k-1} - a_{k-2} &= \frac{1-p}{p}(a_{k-2} - a_{k-3}) &&= \left(\frac{1-p}{p}\right)^{k-2} a_1, \\
1 - a_{k-1} &= \frac{1-p}{p}(a_{k-1} - a_{k-2}) &&= \left(\frac{1-p}{p}\right)^{k-1} a_1.
\end{aligned}
\tag{2.4.4}
$$

By equating the sum of the left sides of these $k-1$ equations with the sum of the right sides, we obtain the relation

$$
1 - a_1 = a_1 \sum_{i=1}^{k-1} \left(\frac{1-p}{p}\right)^i. \tag{2.4.5}
$$

Solution for a Fair Game Suppose first that $p = 1/2$. Then $(1-p)/p = 1$, and it follows from Eq. (2.4.5) that $1 - a_1 = (k-1)a_1$, from which $a_1 = 1/k$. In turn, it follows from the first equation in (2.4.4) that $a_2 = 2/k$, it follows from the second equation in (2.4.4) that $a_3 = 3/k$, and so on. In this way, we obtain the following complete solution when $p = 1/2$:

$$
a_i = \frac{i}{k} \quad \text{for } i = 1, \ldots, k-1. \tag{2.4.6}
$$

Example 2.4.1

The Probability of Winning in a Fair Game. Suppose that $p = 1/2$, in which case the game is equally favorable to both gamblers; and suppose that the initial fortune of gambler A is 98 dollars and the initial fortune of gambler B is just two dollars. In this example, $i = 98$ and $k = 100$. Therefore, it follows from Eq. (2.4.6) that there is a probability of 0.98 that gambler A will win two dollars from gambler B before gambler B wins 98 dollars from gambler A. ◀

Solution for an Unfair Game Suppose now that $p \neq 1/2$. Then Eq. (2.4.5) can be rewritten in the form

$$
1 - a_1 = a_1 \frac{\left(\frac{1-p}{p}\right)^k - \left(\frac{1-p}{p}\right)}{\left(\frac{1-p}{p}\right) - 1}. \tag{2.4.7}
$$

Hence,

$$
a_1 = \frac{\left(\frac{1-p}{p}\right) - 1}{\left(\frac{1-p}{p}\right)^k - 1}. \tag{2.4.8}
$$

Each of the other values of a_i for $i = 2, \dots, k-1$ can now be determined in turn from the equations in (2.4.4). In this way, we obtain the following complete solution:

$$a_i = \frac{\left(\frac{1-p}{p}\right)^i - 1}{\left(\frac{1-p}{p}\right)^k - 1} \quad \text{for } i = 1, \dots, k-1. \tag{2.4.9}$$

Example 2.4.2

The Probability of Winning in an Unfavorable Game. Suppose that $p = 0.4$, in which case the probability that gambler A will win one dollar on any given play is smaller than the probability that he will lose one dollar. Suppose also that the initial fortune of gambler A is 99 dollars and the initial fortune of gambler B is just one dollar. We shall determine the probability that gambler A will win one dollar from gambler B before gambler B wins 99 dollars from gambler A.

In this example, the required probability a_i is given by Eq. (2.4.9), in which $(1-p)/p = 3/2$, $i = 99$, and $k = 100$. Therefore,

$$a_i = \frac{\left(\frac{3}{2}\right)^{99} - 1}{\left(\frac{3}{2}\right)^{100} - 1} \approx \frac{1}{3/2} = \frac{2}{3}.$$

Hence, although the probability that gambler A will win one dollar on any given play is only 0.4, the probability that he will win one dollar before he loses 99 dollars is approximately 2/3. ◀

Summary

We considered a gambler and an opponent who each start with finite amounts of money. The two then play a sequence of games against each other until one of them runs out of money. We were able to calculate the probability that each of them would be the first to run out as a function of the probability of winning the game and of how much money each has at the start.

Exercises

1. Consider the unfavorable game in Example 2.4.2. This time, suppose that the initial fortune of gambler A is i dollars with $i \le 98$. Suppose that the initial fortune of gambler B is $100 - i$ dollars. Show that the probability is greater than 1/2 that gambler A losses i dollars before winning $100 - i$ dollars.

2. Consider the following three different possible conditions in the gambler's ruin problem:

a. The initial fortune of gambler A is two dollars, and the initial fortune of gambler B is one dollar.

b. The initial fortune of gambler A is 20 dollars, and the initial fortune of gambler B is 10 dollars.

c. The initial fortune of gambler A is 200 dollars, and the initial fortune of gambler B is 100 dollars.

Suppose that $p = 1/2$. For which of these three conditions is there the greatest probability that gambler A will win the initial fortune of gambler B before he loses his own initial fortune?

3. Consider again the three different conditions (a), (b), and (c) given in Exercise 2, but suppose now that $p < 1/2$. For which of these three conditions is there the greatest probability that gambler A will win the initial fortune of gambler B before he loses his own initial fortune?

4. Consider again the three different conditions (a), (b), and (c) given in Exercise 2, but suppose now that $p > 1/2$. For which of these three conditions is there the greatest probability that gambler A will win the initial fortune of gambler B before he loses his own initial fortune?

5. Suppose that on each play of a certain game, a person is equally likely to win one dollar or lose one dollar. Suppose also that the person's goal is to win two dollars by playing this game. How large an initial fortune must the person have in order for the probability to be at least 0.99 that she will achieve her goal before she loses her initial fortune?

6. Suppose that on each play of a certain game, a person will either win one dollar with probability 2/3 or lose one dollar with probability 1/3. Suppose also that the person's goal is to win two dollars by playing this game. How large an initial fortune must the person have in order for the probability to be at least 0.99 that he will achieve his goal before he loses his initial fortune?

7. Suppose that on each play of a certain game, a person will either win one dollar with probability 1/3 or lose one dollar with probability 2/3. Suppose also that the person's goal is to win two dollars by playing this game. Show that no matter how large the person's initial fortune might be, the probability that she will achieve her goal before she loses her initial fortune is less than 1/4.

8. Suppose that the probability of a head on any toss of a certain coin is p $(0 < p < 1)$, and suppose that the coin is tossed repeatedly. Let X_n denote the total number of heads that have been obtained on the first n tosses, and let $Y_n = n - X_n$ denote the total number of tails on the first n tosses. Suppose that the tosses are stopped as soon as a number n is reached such that either $X_n = Y_n + 3$ or $Y_n = X_n + 3$. Determine the probability that $X_n = Y_n + 3$ when the tosses are stopped.

9. Suppose that a certain box A contains five balls and another box B contains 10 balls. One of these two boxes is selected at random, and one ball from the selected box is transferred to the other box. If this process of selecting a box at random and transferring one ball from that box to the other box is repeated indefinitely, what is the probability that box A will become empty before box B becomes empty?

2.5 Supplementary Exercises

1. Suppose that A, B, and D are any three events such that $\Pr(A|D) \ge \Pr(B|D)$ and $\Pr(A|D^c) \ge \Pr(B|D^c)$. Prove that $\Pr(A) \ge \Pr(B)$.

2. Suppose that a fair coin is tossed repeatedly and independently until both a head and a tail have appeared at least once. **(a)** Describe the sample space of this experiment. **(b)** What is the probability that exactly three tosses will be required?

3. Suppose that A and B are events such that $\Pr(A) = 1/3$, $\Pr(B) = 1/5$, and $\Pr(A|B) + \Pr(B|A) = 2/3$. Evaluate $\Pr(A^c \cup B^c)$.

4. Suppose that A and B are independent events such that $\Pr(A) = 1/3$ and $\Pr(B) > 0$. What is the value of $\Pr(A \cup B^c|B)$?

5. Suppose that in 10 rolls of a balanced die, the number 6 appeared exactly three times. What is the probability that the first three rolls each yielded the number 6?

6. Suppose that A, B, and D are events such that A and B are independent, $\Pr(A \cap B \cap D) = 0.04$, $\Pr(D|A \cap B) = 0.25$, and $\Pr(B) = 4\Pr(A)$. Evaluate $\Pr(A \cup B)$.

7. Suppose that the events A, B, and C are mutually independent. Under what conditions are A^c, B^c, and C^c mutually independent?

8. Suppose that the events A and B are disjoint and that each has positive probability. Are A and B independent?

9. Suppose that A, B, and C are three events such that A and B are disjoint, A and C are independent, and B and C are independent. Suppose also that $4\Pr(A) = 2\Pr(B) = \Pr(C) > 0$ and $\Pr(A \cup B \cup C) = 5\Pr(A)$. Determine the value of $\Pr(A)$.

10. Suppose that each of two dice is loaded so that when either die is rolled, the probability that the number k will appear is 0.1 for $k = 1, 2, 5$, or 6 and is 0.3 for $k = 3$ or 4. If the two loaded dice are rolled independently, what is the probability that the sum of the two numbers that appear will be 7?

11. Suppose that there is a probability of 1/50 that you will win a certain game. If you play the game 50 times, independently, what is the probability that you will win at least once?

12. Suppose that a balanced die is rolled three times, and let X_i denote the number that appears on the ith roll $(i = 1, 2, 3)$. Evaluate $\Pr(X_1 > X_2 > X_3)$.

13. Three students A, B, and C are enrolled in the same class. Suppose that A attends class 30 percent of the time, B attends class 50 percent of the time, and C attends class 80 percent of the time. If these students attend class independently of each other, what is (a) the probability that at least one of them will be in class on a particular day and (b) the probability that exactly one of them will be in class on a particular day?

14. Consider the World Series of baseball, as described in Exercise 16 of Sec. 2.2. If there is probability p that team A will win any particular game, what is the probability

that it will be necessary to play seven games in order to determine the winner of the Series?

15. Suppose that three red balls and three white balls are thrown at random into three boxes and and that all throws are independent. What is the probability that each box contains one red ball and one white ball?

16. If five balls are thrown at random into n boxes, and all throws are independent, what is the probability that no box contains more than two balls?

17. Bus tickets in a certain city contain four numbers, U, V, W, and X. Each of these numbers is equally likely to be any of the 10 digits 0, 1, . . . , 9, and the four numbers are chosen independently. A bus rider is said to be lucky if $U + V = W + X$. What proportion of the riders are lucky?

18. A certain group has eight members. In January, three members are selected at random to serve on a committee. In February, four members are selected at random and independently of the first selection to serve on another committee. In March, five members are selected at random and independently of the previous two selections to serve on a third committee. Determine the probability that each of the eight members serves on at least one of the three committees.

19. For the conditions of Exercise 18, determine the probability that two particular members A and B will serve together on at least one of the three committees.

20. Suppose that two players A and B take turns rolling a pair of balanced dice and that the winner is the first player who obtains the sum of 7 on a given roll of the two dice. If A rolls first, what is the probability that B will win?

21. Three players A, B, and C take turns tossing a fair coin. Suppose that A tosses the coin first, B tosses second, and C tosses third; and suppose that this cycle is repeated indefinitely until someone wins by being the first player to obtain a head. Determine the probability that each of three players will win.

22. Suppose that a balanced die is rolled repeatedly until the same number appears on two successive rolls, and let X denote the number of rolls that are required. Determine the value of $\Pr(X = x)$, for $x = 2, 3, \ldots$.

23. Suppose that 80 percent of all statisticians are shy, whereas only 15 percent of all economists are shy. Suppose also that 90 percent of the people at a large gathering are economists and the other 10 percent are statisticians. If you meet a shy person at random at the gathering, what is the probability that the person is a statistician?

24. Dreamboat cars are produced at three different factories A, B, and C. Factory A produces 20 percent of the total output of Dreamboats, B produces 50 percent, and C produces 30 percent. However, 5 percent of the cars produced at A are lemons, 2 percent of those produced at B are lemons, and 10 percent of those produced at C are lemons. If you buy a Dreamboat and it turns out to be a lemon, what is the probability that it was produced at factory A?

25. Suppose that 30 percent of the bottles produced in a certain plant are defective. If a bottle is defective, the probability is 0.9 that an inspector will notice it and remove it from the filling line. If a bottle is not defective, the probability is 0.2 that the inspector will think that it is defective and remove it from the filling line.

a. If a bottle is removed from the filling line, what is the probability that it is defective?

b. If a customer buys a bottle that has not been removed from the filling line, what is the probability that it is defective?

26. Suppose that a fair coin is tossed until a head is obtained and that this entire experiment is then performed independently a second time. What is the probability that the second experiment requires more tosses than the first experiment?

27. Suppose that a family has exactly n children ($n \geq 2$). Assume that the probability that any child will be a girl is 1/2 and that all births are independent. Given that the family has at least one girl, determine the probability that the family has at least one boy.

28. Suppose that a fair coin is tossed independently n times. Determine the probability of obtaining exactly $n - 1$ heads, given **(a)** that at least $n - 2$ heads are obtained and **(b)** that heads are obtained on the first $n - 2$ tosses.

29. Suppose that 13 cards are selected at random from a regular deck of 52 playing cards.

a. If it is known that at least one ace has been selected, what is the probability that at least two aces have been selected?

b. If it is known that the ace of hearts has been selected, what is the probability that at least two aces have been selected?

30. Suppose that n letters are placed at random in n envelopes, as in the matching problem of Sec. 1.10, and let q_n denote the probability that no letter is placed in the correct envelope. Show that the probability that exactly one letter is placed in the correct envelope is q_{n-1}.

31. Consider again the conditions of Exercise 30. Show that the probability that exactly two letters are placed in the correct envelopes is $(1/2)q_{n-2}$.

32. Consider again the conditions of Exercise 7 of Sec. 2.2. If exactly one of the two students A and B is in class on a given day, what is the probability that it is A?

33. Consider again the conditions of Exercise 2 of Sec. 1.10. If a family selected at random from the city

subscribes to exactly one of the three newspapers A, B, and C, what is the probability that it is A?

34. Three prisoners A, B, and C on death row know that exactly two of them are going to be executed, but they do not know which two. Prisoner A knows that the jailer will not tell him whether or not he is going to be executed. He therefore asks the jailer to tell him the name of one prisoner other than A himself who will be executed. The jailer responds that B will be executed. Upon receiving this response, Prisoner A reasons as follows: Before he spoke to the jailer, the probability was 2/3 that he would be one of the two prisoners executed. After speaking to the jailer, he knows that either he or prisoner C will be the other one to be executed. Hence, the probability that he will be executed is now only 1/2. Thus, merely by asking the jailer his question, the prisoner reduced the probability that he would be executed from 2/3 to 1/2, because he could go through exactly this same reasoning regardless of which answer the jailer gave. Discuss what is wrong with prisoner A's reasoning.

35. Suppose that each of two gamblers A and B has an initial fortune of 50 dollars, and that there is probability p that gambler A will win on any single play of a game against gambler B. Also, suppose either that one gambler can win one dollar from the other on each play of the game or that they can double the stakes and one can win two dollars from the other on each play of the game. Under which of these two conditions does A have the greater probability of winning the initial fortune of B before losing her own for each of the following conditions: **(a)** $p < 1/2$; **(b)** $p > 1/2$; **(c)** $p = 1/2$?

36. A sequence of n job candidates is prepared to interview for a job. We would like to hire the best candidate, but we have no information to distinguish the candidates before we interview them. We assume that the best candidate is equally likely to be each of the n candidates in the sequence before the interviews start. After the interviews start, we are able to rank those candidates we have seen, but we have no information about where the remaining candidates rank relative to those we have seen. After each interview, it is required that either we hire the current candidate immediately and stop the interviews, or we must let the current candidate go and we never can call them back. We choose to interview as follows: We select a number $0 \le r < n$ and we interview the first r candidates without any intention of hiring them. Starting with the next candidate $r + 1$, we continue interviewing until the current candidate is the best we have seen so far. We then stop and hire the current candidate. If none of the candidates from $r + 1$ to n is the best, we just hire candidate n. We would like to compute the probability that we hire the best candidate and we would like to choose r to make this probability as large as possible. Let A be the event that we hire the best candidate, and let B_i be the event that the best candidate is in position i in the sequence of interviews.

a. Let $i > r$. Find the probability that the candidate who is relatively the best among the first i interviewed appears in the first r interviews.

b. Prove that $\Pr(A|B_i) = 0$ for $i \le r$ and $\Pr(A|B_i) = r/(i-1)$ for $i > r$.

c. For fixed r, let p_r be the probability of A using that value of r. Prove that $p_r = (r/n)\sum_{i=r+1}^{n}(i-1)^{-1}$.

d. Let $q_r = p_r - p_{r-1}$ for $r = 1, \dots, n-1$, and prove that q_r is a strictly decreasing function of r.

e. Show that a value of r that maximizes p_r is the last r such that $q_r > 0$. (*Hint:* Write $p_r = p_0 + q_1 + \cdots + q_r$ for $r > 0$.)

f. For $n = 10$, find the value of r that maximizes p_r, and find the corresponding p_r value.

Chapter 3

Random Variables and Distributions

3.1 Random Variables and Discrete Distributions

A random variable is a real-valued function defined on a sample space. Random variables are the main tools used for modeling unknown quantities in statistical analyses. For each random variable X and each set C of real numbers, we could calculate the probability that X takes its value in C. The collection of all of these probabilities is the distribution of X. There are two major classes of distributions and random variables: discrete (this section) and continuous (Sec. 3.2). Discrete distributions are those that assign positive probability to at most countably many different values. A discrete distribution can be characterized by its probability function (p.f.), which specifies the probability that the random variable takes each of the different possible values. A random variable with a discrete distribution will be called a discrete random variable.

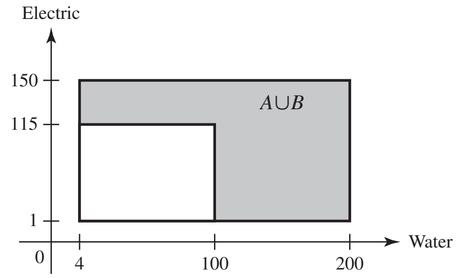
Definition of a Random Variable

Example 3.1.1 Tossing a Coin. Consider an experiment in which a fair coin is tossed 10 times. In this experiment, the sample space S can be regarded as the set of outcomes consisting of the 2^{10} different sequences of 10 heads and/or tails that are possible. We might be interested in the number of heads in the observed outcome. We can let X stand for the real-valued function defined on S that counts the number of heads in each outcome. For example, if s is the sequence HHTTTHTTTH, then $X(s) = 4$. For each possible sequence s consisting of 10 heads and/or tails, the value $X(s)$ equals the number of heads in the sequence. The possible values for the function X are 0, 1, . . . , 10. ◄

Definition 3.1.1 Random Variable. Let S be the sample space for an experiment. A real-valued function that is defined on S is called a *random variable*.

For example, in Example 3.1.1, the number X of heads in the 10 tosses is a random variable. Another random variable in that example is $Y = 10 - X$, the number of tails.

Figure 3.1 The event that at least one utility demand is high in Example 3.1.3.



Example 3.1.2 Measuring a Person's Height. Consider an experiment in which a person is selected at random from some population and her height in inches is measured. This height is a random variable. ◀

Example 3.1.3 Demands for Utilities. Consider the contractor in Example 1.5.4 on page 19 who is concerned about the demands for water and electricity in a new office complex. The sample space was pictured in Fig. 1.5 on page 12, and it consists of a collection of points of the form (x, y), where x is the demand for water and y is the demand for electricity. That is, each point $s \in S$ is a pair $s = (x, y)$. One random variable that is of interest in this problem is the demand for water. This can be expressed as $X(s) = x$ when $s = (x, y)$. The possible values of X are the numbers in the interval [4, 200]. Another interesting random variable is Y, equal to the electricity demand, which can be expressed as $Y(s) = y$ when $s = (x, y)$. The possible values of Y are the numbers in the interval [1, 150]. A third possible random variable Z is an indicator of whether or not at least one demand is high. Let A and B be the two events described in Example 1.5.4. That is, A is the event that water demand is at least 100, and B is the event that electric demand is at least 115. Define

$$Z(s) = \begin{cases} 1 & \text{if } s \in A \cup B, \\ 0 & \text{if } s \notin A \cup B. \end{cases}$$

The possible values of Z are the numbers 0 and 1. The event $A \cup B$ is indicated in Fig. 3.1. ◀

The Distribution of a Random Variable

When a probability measure has been specified on the sample space of an experiment, we can determine probabilities associated with the possible values of each random variable X. Let C be a subset of the real line such that $\{X \in C\}$ is an event, and let $\Pr(X \in C)$ denote the probability that the value of X will belong to the subset C. Then $\Pr(X \in C)$ is equal to the probability that the outcome s of the experiment will be such that $X(s) \in C$. In symbols,

$$\Pr(X \in C) = \Pr(\{s: X(s) \in C\}). \tag{3.1.1}$$

Definition 3.1.2 Distribution. Let X be a random variable. The *distribution* of X is the collection of all probabilities of the form $\Pr(X \in C)$ for all sets C of real numbers such that $\{X \in C\}$ is an event.

It is a straightforward consequence of the definition of the distribution of X that this distribution is itself a probability measure on the set of real numbers. The set

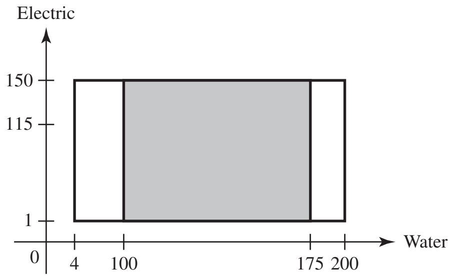


Figure 3.2 The event that water demand is between 50 and 175 in Example 3.1.5.

$\{X \in C\}$ will be an event for every set C of real numbers that most readers will be able to imagine.

Example 3.1.4

Tossing a Coin. Consider again an experiment in which a fair coin is tossed 10 times, and let X be the number of heads that are obtained. In this experiment, the possible values of X are 0, 1, 2, . . . , 10. For each x, $\Pr(X = x)$ is the sum of the probabilities of all of the outcomes in the event $\{X = x\}$. Because the coin is fair, each outcome has the same probability $1/2^{10}$, and we need only count how many outcomes s have $X(s) = x$. We know that $X(s) = x$ if and only if exactly x of the 10 tosses are H. Hence, the number of outcomes s with $X(s) = x$ is the same as the number of subsets of size x (to be the heads) that can be chosen from the 10 tosses, namely, $\binom{10}{x}$, according to Definitions 1.8.1 and 1.8.2. Hence,

$$\Pr(X = x) = \binom{10}{x}\frac{1}{2^{10}} \quad \text{for } x = 0, 1, 2, \ldots, 10. \qquad ◀$$

Example 3.1.5

Demands for Utilities. In Example 1.5.4, we actually calculated some features of the distributions of the three random variables X, Y, and Z defined in Example 3.1.3. For example, the event A, defined as the event that water demand is at least 100, can be expressed as $A = \{X \geq 100\}$, and $\Pr(A) = 0.5102$. This means that $\Pr(X \geq 100) = 0.5102$. The distribution of X consists of all probabilities of the form $\Pr(X \in C)$ for all sets C such that $\{X \in C\}$ is an event. These can all be calculated in a manner similar to the calculation of $\Pr(A)$ in Example 1.5.4. In particular, if C is a subinterval of the interval [4, 200], then

$$\Pr(X \in C) = \frac{(150 - 1) \times (\text{length of interval } C)}{29{,}204}. \tag{3.1.2}$$

For example, if C is the interval [50,175], then its length is 125, and $\Pr(X \in C) = 149 \times 125/29{,}204 = 0.6378$. The subset of the sample space whose probability was just calculated is drawn in Fig. 3.2. ◀

The general definition of distribution in Definition 3.1.2 is awkward, and it will be useful to find alternative ways to specify the distributions of random variables. In the remainder of this section, we shall introduce a few such alternatives.

Discrete Distributions

Definition 3.1.3

Discrete Distribution/Random Variable. We say that a random variable X has a *discrete distribution* or that X is a *discrete random variable* if X can take only a finite number k of different values $x_1, \ldots, x_k$ or, at most, an infinite sequence of different values $x_1, x_2, \ldots$.

Random variables that can take every value in an interval are said to have *continuous distributions* and are discussed in Sec. 3.2.

Definition 3.1.4 Probability Function/p.f./Support. If a random variable X has a discrete distribution, the *probability function* (abbreviated *p.f.*) of X is defined as the function f such that for every real number x,

$$f(x) = \Pr(X = x).$$

The closure of the set $\{x : f(x) > 0\}$ is called the *support of (the distribution of) X*.

Some authors refer to the probability function as the *probability mass function,* or p.m.f. We will not use that term again in this text.

Example 3.1.6 Demands for Utilities. The random variable Z in Example 3.1.3 equals 1 if at least one of the utility demands is high, and $Z = 0$ if neither demand is high. Since Z takes only two different values, it has a discrete distribution. Note that $\{s : Z(s) = 1\} = A \cup B$, where A and B are defined in Example 1.5.4. We calculated $\Pr(A \cup B) = 0.65253$ in Example 1.5.4. If Z has p.f. f, then

$$f(z) = \begin{cases} 0.65253 & \text{if } z = 1, \\ 0.34747 & \text{if } z = 0, \\ 0 & \text{otherwise.} \end{cases}$$

The support of Z is the set $\{0, 1\}$, which has only two elements. ◀

Example 3.1.7 Tossing a Coin. The random variable X in Example 3.1.4 has only 11 different possible values. Its p.f. f is given at the end of that example for the values $x = 0, \dots, 10$ that constitute the support of X; $f(x) = 0$ for all other values of x. ◀

Here are some simple facts about probability functions

Theorem 3.1.1 Let X be a discrete random variable with p.f. f. If x is not one of the possible values of X, then $f(x) = 0$. Also, if the sequence $x_1, x_2, \dots$ includes all the possible values of X, then $\sum_{i=1}^{\infty} f(x_i) = 1$. ■

A typical p.f. is sketched in Fig. 3.3, in which each vertical segment represents the value of $f(x)$ corresponding to a possible value x. The sum of the heights of the vertical segments in Fig. 3.3 must be 1.

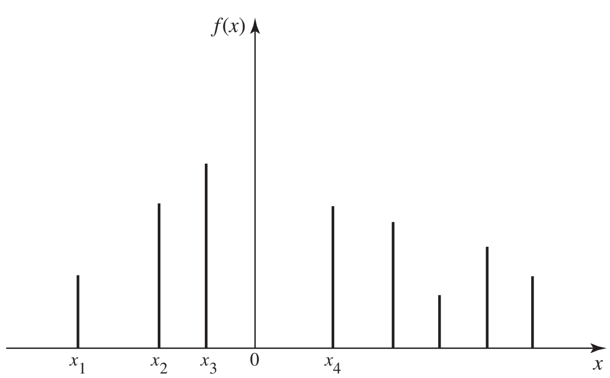


Figure 3.3 An example of a p.f.

Theorem 3.1.2 shows that the p.f. of a discrete random variable characterizes its distribution, and it allows us to dispense with the general definition of distribution when we are discussing discrete random variables.

Theorem 3.1.2 If X has a discrete distribution, the probability of each subset C of the real line can be determined from the relation

$$\Pr(X \in C) = \sum_{x_i \in C} f(x_i). \qquad \blacksquare$$

Some random variables have distributions that appear so frequently that the distributions are given names. The random variable Z in Example 3.1.6 is one such.

Definition 3.1.5 Bernoulli Distribution/Random Variable. A random variable Z that takes only two values 0 and 1 with $\Pr(Z = 1) = p$ has the *Bernoulli distribution with parameter* p. We also say that Z is a *Bernoulli random variable with parameter* p.

The Z in Example 3.1.6 has the Bernoulli distribution with parameter 0.65252. It is easy to see that the name of each Bernoulli distribution is enough to allow us to compute the p.f., which, in turn, allows us to characterize its distribution.

We conclude this section with illustrations of two additional families of discrete distributions that arise often enough to have names.

Uniform Distributions on Integers

Example 3.1.8 Daily Numbers. A popular state lottery game requires participants to select a three-digit number (leading 0s allowed). Then three balls, each with one digit, are chosen at random from well-mixed bowls. The sample space here consists of all triples (i_1, i_2, i_3) where $i_j \in \{0, \ldots, 9\}$ for $j = 1, 2, 3$. If $s = (i_1, i_2, i_3)$, define $X(s) = 100i_1 + 10i_2 + i_3$. For example, $X(0, 1, 5) = 15$. It is easy to check that $\Pr(X = x) = 0.001$ for each integer $x \in \{0, 1, \ldots, 999\}$. ◀

Definition 3.1.6 Uniform Distribution on Integers. Let $a \le b$ be integers. Suppose that the value of a random variable X is equally likely to be each of the integers $a, \ldots, b$. Then we say that X has the *uniform distribution on the integers* $a, \ldots, b$.

The X in Example 3.1.8 has the uniform distribution on the integers $0, 1, \ldots, 999$. A uniform distribution on a set of k integers has probability $1/k$ on each integer. If $b > a$, there are $b - a + 1$ integers from a to b including a and b. The next result follows immediately from what we have just seen, and it illustrates how the name of the distribution characterizes the distribution.

Theorem 3.1.3 If X has the uniform distribution on the integers $a, \ldots, b$, the p.f. of X is

$$f(x) = \begin{cases} \dfrac{1}{b - a + 1} & \text{for } x = a, \ldots, b, \\ 0 & \text{otherwise.} \end{cases} \qquad \blacksquare$$

The uniform distribution on the integers $a, \ldots, b$ represents the outcome of an experiment that is often described by saying that one of the integers $a, \ldots, b$ is *chosen at random*. In this context, the phrase "at random" means that each of the $b - a + 1$ integers is equally likely to be chosen. In this same sense, it is not possible to choose an integer at random from the set of *all* positive integers, because it is not possible

to assign the same probability to every one of the positive integers and still make the sum of these probabilities equal to 1. In other words, a uniform distribution cannot be assigned to an infinite sequence of possible values, but such a distribution can be assigned to any finite sequence.

Note: Random Variables Can Have the Same Distribution without Being the Same Random Variable. Consider two consecutive daily number draws as in Example 3.1.8. The sample space consists of all 6-tuples $(i_1, \ldots, i_6)$, where the first three coordinates are the numbers drawn on the first day and the last three are the numbers drawn on the second day (all in the order drawn). If $s = (i_1, \ldots, i_6)$, let $X_1(s) = 100i_1 + 10i_2 + i_3$ and let $X_2(s) = 100i_4 + 10i_5 + i_6$. It is easy to see that X_1 and X_2 are different functions of s and are not the same random variable. Indeed, there is only a small probability that they will take the same value. But they have the same distribution because they assume the same values with the same probabilities. If a businessman has 1000 customers numbered $0, \ldots, 999$, and he selects one at random and records the number Y, the distribution of Y will be the same as the distribution of X_1 and of X_2, but Y is not like X_1 or X_2 in any other way.

Binomial Distributions

Example 3.1.9 Defective Parts. Consider again Example 2.2.5 from page 69. In that example, a machine produces a defective item with probability p $(0 < p < 1)$ and produces a nondefective item with probability $1 - p$. We assumed that the events that the different items were defective were mutually independent. Suppose that the experiment consists of examining n of these items. Each outcome of this experiment will consist of a list of which items are defective and which are not, in the order examined. For example, we can let 0 stand for a nondefective item and 1 stand for a defective item. Then each outcome is a string of n digits, each of which is 0 or 1. To be specific, if, say, $n = 6$, then some of the possible outcomes are

$$010010,\ 100100,\ 000011,\ 110000,\ 100001,\ 000000,\ \text{etc.} \tag{3.1.3}$$

We will let X denote the number of these items that are defective. Then the random variable X will have a discrete distribution, and the possible values of X will be $0, 1, 2, \ldots, n$. For example, the first four outcomes listed in Eq. (3.1.3) all have $X(s) = 2$. The last outcome listed has $X(s) = 0$. ◄

Example 3.1.9 is a generalization of Example 2.2.5 with n items inspected rather than just six, and rewritten in the notation of random variables. For $x = 0, 1, \ldots, n$, the probability of obtaining each particular ordered sequence of n items containing exactly x defectives and $n - x$ nondefectives is $p^x(1-p)^{n-x}$, just as it was in Example 2.2.5. Since there are $\binom{n}{x}$ different ordered sequences of this type, it follows that

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Therefore, the p.f. of X will be as follows:

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, \ldots, n, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1.4}$$

Definition 3.1.7 Binomial Distribution/Random Variable. The discrete distribution represented by the p.f. in (3.1.4) is called the *binomial distribution with parameters n and p*. A random

variable with this distribution is said to be a *binomial random variable with parameters n and p*.

The reader should be able to verify that the random variable X in Example 3.1.4, the number of heads in a sequence of 10 independent tosses of a fair coin, has the binomial distribution with parameters 10 and 1/2.

Since the name of each binomial distribution is sufficient to construct its p.f., it follows that the name is enough to identify the distribution. The name of each distribution includes the two parameters. The binomial distributions are very important in probability and statistics and will be discussed further in later chapters of this book.

A short table of values of certain binomial distributions is given at the end of this book. It can be found from this table, for example, that if X has the binomial distribution with parameters $n = 10$ and $p = 0.2$, then $\Pr(X = 5) = 0.0264$ and $\Pr(X \geq 5) = 0.0328$.

As another example, suppose that a clinical trial is being run. Suppose that the probability that a patient recovers from her symptoms during the trial is p and that the probability is $1 - p$ that the patient does not recover. Let Y denote the number of patients who recover out of n independent patients in the trial. Then the distribution of Y is also binomial with parameters n and p. Indeed, consider a general experiment that consists of observing n independent repititions (trials) with only two possible results for each trial. For convenience, call the two possible results "success" and "failure." Then the distribution of the number of trials that result in success will be binomial with parameters n and p, where p is the probability of success on each trial.

Note: Names of Distributions. In this section, we gave names to several families of distributions. The name of each distribution includes any numerical parameters that are part of the definition. For example, the random variable X in Example 3.1.4 has the binomial distribution with parameters 10 and 1/2. It is a correct statement to say that X has a binomial distribution or that X has a discrete distribution, but such statements are only partial descriptions of the distribution of X. Such statements are *not* sufficient to name the distribution of X, and hence they are not sufficient as answers to the question "What is the distribution of X?" The same considerations apply to all of the named distributions that we introduce elsewhere in the book. When attempting to specify the distribution of a random variable by giving its name, one must give the full name, including the values of any parameters. Only the full name is sufficient for determining the distribution.

Summary

A random variable is a real-valued function defined on a sample space. The distribution of a random variable X is the collection of all probabilities $\Pr(X \in C)$ for all subsets C of the real numbers such that $\{X \in C\}$ is an event. A random variable X is discrete if there are at most countably many possible values for X. In this case, the distribution of X can be characterized by the probability function (p.f.) of X, namely, $f(x) = \Pr(X = x)$ for x in the set of possible values. Some distributions are so famous that they have names. One collection of such named distributions is the collection of uniform distributions on finite sets of integers. A more famous collection is the collection of binomial distributions whose parameters are n and p, where n is a positive integer and $0 < p < 1$, having p.f. (3.1.4). The binomial distribution with parameters $n = 1$ and p is also called the Bernoulli distribution with parameter p. The names of these distributions also characterize the distributions.

Exercises

1. Suppose that a random variable X has the uniform distribution on the integers $10, \dots, 20$. Find the probability that X is even.

2. Suppose that a random variable X has a discrete distribution with the following p.f.:

$$f(x) = \begin{cases} cx & \text{for } x = 1, \dots, 5, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the value of the constant c.

3. Suppose that two balanced dice are rolled, and let X denote the absolute value of the difference between the two numbers that appear. Determine and sketch the p.f. of X.

4. Suppose that a fair coin is tossed 10 times independently. Determine the p.f. of the number of heads that will be obtained.

5. Suppose that a box contains seven red balls and three blue balls. If five balls are selected at random, without replacement, determine the p.f. of the number of red balls that will be obtained.

6. Suppose that a random variable X has the binomial distribution with parameters $n = 15$ and $p = 0.5$. Find $\Pr(X < 6)$.

7. Suppose that a random variable X has the binomial distribution with parameters $n = 8$ and $p = 0.7$. Find $\Pr(X \geq 5)$ by using the table given at the end of this book. *Hint:* Use the fact that $\Pr(X \geq 5) = \Pr(Y \leq 3)$, where Y has the binomial distribution with parameters $n = 8$ and $p = 0.3$.

8. If 10 percent of the balls in a certain box are red, and if 20 balls are selected from the box at random, with replacement, what is the probability that more than three red balls will be obtained?

9. Suppose that a random variable X has a discrete distribution with the following p.f.:

$$f(x) = \begin{cases} \frac{c}{2^x} & \text{for } x = 0, 1, 2, \dots, \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of the constant c.

10. A civil engineer is studying a left-turn lane that is long enough to hold seven cars. Let X be the number of cars in the lane at the end of a randomly chosen red light. The engineer believes that the probability that $X = x$ is proportional to $(x + 1)(8 - x)$ for $x = 0, \dots, 7$ (the possible values of X).

a. Find the p.f. of X.

b. Find the probability that X will be at least 5.

11. Show that there does not exist any number c such that the following function would be a p.f.:

$$f(x) = \begin{cases} \frac{c}{x} & \text{for } x = 1, 2, \dots, \\ 0 & \text{otherwise.} \end{cases}$$

3.2 Continuous Distributions

Next, we focus on random variables that can assume every value in an interval (bounded or unbounded). If a random variable X has associated with it a function f such that the integral of f over each interval gives the probability that X is in the interval, then we call f the probability density function (p.d.f.) of X and we say that X has a continuous distribution.

The Probability Density Function

Example 3.2.1 Demands for Utilities. In Example 3.1.5, we determined the distribution of the demand for water, X. From Fig. 3.2, we see that the smallest possible value of X is 4 and the largest is 200. For each interval $C = [c_0, c_1] \subset [4, 200]$, Eq. (3.1.2) says that

$$\Pr(c_0 \leq X \leq c_1) = \frac{149(c_1 - c_0)}{29204} = \frac{c_1 - c_0}{196} = \int_{c_0}^{c_1} \frac{1}{196} dx.$$

So, if we define

$$f(x) = \begin{cases} \dfrac{1}{196} & \text{if } 4 \le x \le 200, \\ 0 & \text{otherwise,} \end{cases} \tag{3.2.1}$$

we have that

$$\Pr(c_0 \le X \le c_1) = \int_{c_0}^{c_1} f(x)dx. \tag{3.2.2}$$

Because we defined $f(x)$ to be 0 for x outside of the interval [4, 200], we see that Eq. (3.2.2) holds for all $c_0 \le c_1$, even if $c_0 = -\infty$ and/or $c_1 = \infty$. ◀

The water demand X in Example 3.2.1 is an example of the following.

Definition 3.2.1 Continuous Distribution/Random Variable. We say that a random variable X has a *continuous distribution* or that X is a *continuous random variable* if there exists a nonnegative function f, defined on the real line, such that for every interval of real numbers (bounded or unbounded), the probability that X takes a value in the interval is the integral of f over the interval.

For example, in the situation described in Definition 3.2.1, for each bounded closed interval $[a, b]$,

$$\Pr(a \le X \le b) = \int_a^b f(x)\, dx. \tag{3.2.3}$$

Similarly, $\Pr(X \ge a) = \int_a^\infty f(x)\, dx$ and $\Pr(X \le b) = \int_{-\infty}^b f(x)\, dx$.

We see that the function f characterizes the distribution of a continuous random variable in much the same way that the probability function characterizes the distribution of a discrete random variable. For this reason, the function f plays an important role, and hence we give it a name.

Definition 3.2.2 Probability Density Function/p.d.f./Support. If X has a continuous distribution, the function f described in Definition 3.2.1 is called the *probability density function* (abbreviated *p.d.f.*) of X. The closure of the set $\{x : f(x) > 0\}$ is called the *support of (the distribution of) X*.

Example 3.2.1 demonstrates that the water demand X has p.d.f. given by (3.2.1).

Every p.d.f. f must satisfy the following two requirements:

$$f(x) \ge 0, \quad \text{for all x}, \tag{3.2.4}$$

and

$$\int_{-\infty}^{\infty} f(x)\, dx = 1. \tag{3.2.5}$$

A typical p.d.f. is sketched in Fig. 3.4. In that figure, the total area under the curve must be 1, and the value of $\Pr(a \le X \le b)$ is equal to the area of the shaded region.

Note: Continuous Distributions Assign Probability 0 to Individual Values. The integral in Eq. (3.2.3) also equals $\Pr(a < X \le b)$ as well as $\Pr(a < X < b)$ and $\Pr(a \le X < b)$. Hence, it follows from the definition of continuous distributions that, if X has a continuous distribution, $\Pr(X = a) = 0$ for each number a. As we noted on page 20, the fact that $\Pr(X = a) = 0$ does not imply that $X = a$ is impossible. If it did,

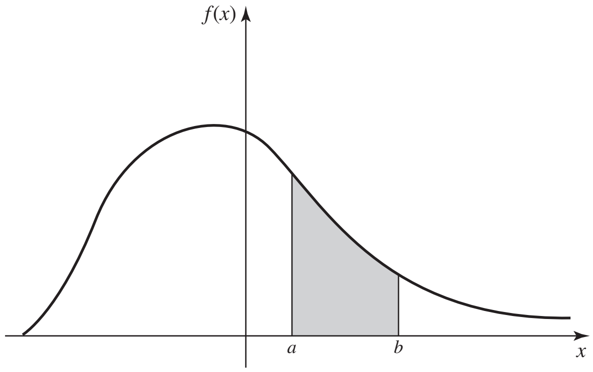


Figure 3.4 An example of a p.d.f.

all values of X would be impossible and X couldn't assume any value. What happens is that the probability in the distribution of X is spread so thinly that we can only see it on sets like nondegenerate intervals. It is much the same as the fact that lines have 0 area in two dimensions, but that does not mean that lines are not there. The two vertical lines indicated under the curve in Fig. 3.4 have 0 area, and this signifies that $\Pr(X = a) = \Pr(X = b) = 0$. However, for each $\epsilon > 0$ and each a such that $f(a) > 0$, $\Pr(a - \epsilon \le X \le a + \epsilon) \approx 2\epsilon f(a) > 0$.

Nonuniqueness of the p.d.f.

If a random variable X has a continuous distribution, then $\Pr(X = x) = 0$ for every individual value x. Because of this property, the values of each p.d.f. can be changed at a finite number of points, or even at certain infinite sequences of points, without changing the value of the integral of the p.d.f. over any subset A. In other words, the values of the p.d.f. of a random variable X can be changed arbitrarily at many points without affecting any probabilities involving X, that is, without affecting the probability distribution of X. At exactly which sets of points we can change a p.d.f. depends on subtle features of the definition of the Riemann integral. We shall not deal with this issue in this text, and we shall only contemplate changes to p.d.f.'s at finitely many points.

To the extent just described, the p.d.f. of a random variable is not unique. In many problems, however, there will be one version of the p.d.f. that is more natural than any other because for this version the p.d.f. will, wherever possible, be continuous on the real line. For example, the p.d.f. sketched in Fig. 3.4 is a continuous function over the entire real line. This p.d.f. could be changed arbitrarily at a few points without affecting the probability distribution that it represents, but these changes would introduce discontinuities into the p.d.f. without introducing any apparent advantages.

Throughout most of this book, we shall adopt the following practice: If a random variable X has a continuous distribution, we shall give only one version of the p.d.f. of X and we shall refer to that version as *the* p.d.f. of X, just as though it had been uniquely determined. It should be remembered, however, that there is some freedom in the selection of the particular version of the p.d.f. that is used to represent each continuous distribution. The most common place where such freedom will arise is in cases like Eq. (3.2.1) where the p.d.f. is required to have discontinuities. Without making the function f any less continuous, we could have defined the p.d.f. in that example so that $f(4) = f(200) = 0$ instead of $f(4) = f(200) = 1/196$. Both of these choices lead to the same calculations of all probabilities associated with X, and they

are both equally valid. Because the support of a continuous distribution is the closure of the set where the p.d.f. is strictly positive, it can be shown that the support is unique. A sensible approach would then be to choose the version of the p.d.f. that was strictly positive on the support whenever possible.

The reader should note that "continuous distribution" is *not* the name of a distribution, just as "discrete distribution" is not the name of a distribution. There are many distributions that are discrete and many that are continuous. Some distributions of each type have names that we either have introduced or will introduce later.

We shall now present several examples of continuous distributions and their p.d.f.'s.

Uniform Distributions on Intervals

Example 3.2.2 Temperature Forecasts. Television weather forecasters announce high and low temperature forecasts as integer numbers of degrees. These forecasts, however, are the results of very sophisticated weather models that provide more precise forecasts that the television personalities round to the nearest integer for simplicity. Suppose that the forecaster announces a high temperature of y. If we wanted to know what temperature X the weather models actually produced, it might be safe to assume that X was equally likely to be any number in the interval from $y - 1/2$ to $y + 1/2$. ◄

The distribution of X in Example 3.2.2 is a special case of the following.

Definition 3.2.3 Uniform Distribution on an Interval. Let a and b be two given real numbers such that $a < b$. Let X be a random variable such that it is known that $a \le X \le b$ and, for every subinterval of $[a, b]$, the probability that X will belong to that subinterval is proportional to the length of that subinterval. We then say that the random variable X has the *uniform distribution on the interval* $[a, b]$.

A random variable X with the uniform distribution on the interval $[a, b]$ represents the outcome of an experiment that is often described by saying that a point is chosen *at random* from the interval $[a, b]$. In this context, the phrase "at random" means that the point is just as likely to be chosen from any particular part of the interval as from any other part of the same length.

Theorem 3.2.1 Uniform Distribution p.d.f. If X has the uniform distribution on an interval $[a, b]$, then the p.d.f. of X is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{otherwise.} \end{cases} \tag{3.2.6}$$

Proof X must take a value in the interval $[a, b]$. Hence, the p.d.f. $f(x)$ of X must be 0 outside of $[a, b]$. Furthermore, since any particular subinterval of $[a, b]$ having a given length is as likely to contain X as is any other subinterval having the same length, regardless of the location of the particular subinterval in $[a, b]$, it follows that $f(x)$ must be constant throughout $[a, b]$, and that interval is then the support of the distribution. Also,

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_a^b f(x)\,dx = 1. \tag{3.2.7}$$

Therefore, the constant value of $f(x)$ throughout $[a, b]$ must be $1/(b-a)$, and the p.d.f. of X must be (3.2.6). ■

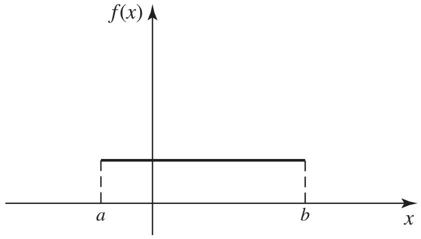


Figure 3.5 The p.d.f. for the uniform distribution on the interval $[a, b]$.

Th p.d.f. (3.2.6) is sketched in Fig. 3.5. As an example, the random variable X (demand for water) in Example 3.2.1 has the uniform distribution on the interval [4, 200].

Note: Density Is Not Probability. The reader should note that the p.d.f. in (3.2.6) can be greater than 1, particularly if $b - a < 1$. Indeed, p.d.f.'s can be unbounded, as we shall see in Example 3.2.6. The p.d.f. of X, $f(x)$, itself does not equal the probability that X is near x. The integral of f over values near x gives the probability that X is near x, and the integral is never greater than 1.

It is seen from Eq. (3.2.6) that the p.d.f. representing a uniform distribution on a given interval is constant over that interval, and the constant value of the p.d.f. is the reciprocal of the length of the interval. It is not possible to define a uniform distribution over an unbounded interval, because the length of such an interval is infinite.

Consider again the uniform distribution on the interval $[a, b]$. Since the probability is 0 that one of the endpoints a or b will be chosen, it is irrelevant whether the distribution is regarded as a uniform distribution on the *closed* interval $a \le x \le b$, or as a uniform distribution on the *open* interval $a < x < b$, or as a uniform distribution on the half-open and half-closed interval $(a, b]$ in which one endpoint is included and the other endpoint is excluded.

For example, if a random variable X has the uniform distribution on the interval $[-1, 4]$, then the p.d.f. of X is

$$f(x) = \begin{cases} 1/5 & \text{for } -1 \le x \le 4, \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore,

$$\Pr(0 \le X < 2) = \int_0^2 f(x)\,dx = \frac{2}{5}.$$

Notice that we defined the p.d.f. of X to be strictly positive on the closed interval $[-1, 4]$ and 0 outside of this closed interval. It would have been just as sensible to define the p.d.f. to be strictly positive on the open interval $(-1, 4)$ and 0 outside of this open interval. The probability distribution would be the same either way, including the calculation of $\Pr(0 \le X < 2)$ that we just performed. After this, when there are several equally sensible choices for how to define a p.d.f., we will simply choose one of them without making any note of the other choices.

Other Continuous Distributions

Example 3.2.3

Incompletely Specified p.d.f. Suppose that the p.d.f. of a certain random variable X has the following form:

$$f(x) = \begin{cases} cx & \text{for } 0 < x < 4, \\ 0 & \text{otherwise,} \end{cases}$$

where c is a given constant. We shall determine the value of c.

For every p.d.f., it must be true that $\int_{-\infty}^{\infty} f(x) = 1$. Therefore, in this example,

$$\int_0^4 cx\,dx = 8c = 1.$$

Hence, $c = 1/8$. ◀

Note: Calculating Normalizing Constants. The calculation in Example 3.2.3 illustrates an important point that simplifies many statistical results. The p.d.f. of X was specified without explicitly giving the value of the constant c. However, we were able to figure out what was the value of c by using the fact that the integral of a p.d.f. must be 1. It will often happen, especially in Chapter 8 where we find sampling distributions of summaries of observed data, that we can determine the p.d.f. of a random variable except for a constant factor. That constant factor must be the unique value such that the integral of the p.d.f. is 1, even if we cannot calculate it directly.

Example 3.2.4

Calculating Probabilities from a p.d.f. Suppose that the p.d.f. of X is as in Example 3.2.3, namely,

$$f(x) = \begin{cases} \dfrac{x}{8} & \text{for } 0 < x < 4, \\ 0 & \text{otherwise.} \end{cases}$$

We shall now determine the values of $\Pr(1 \le X \le 2)$ and $\Pr(X > 2)$. Apply Eq. (3.2.3) to get

$$\Pr(1 \le X \le 2) = \int_1^2 \frac{1}{8} x\,dx = \frac{3}{16}$$

and

$$\Pr(X > 2) = \int_2^4 \frac{1}{8} x\,dx = \frac{3}{4}.$$ ◀

Example 3.2.5

Unbounded Random Variables. It is often convenient and useful to represent a continuous distribution by a p.d.f. that is positive over an unbounded interval of the real line. For example, in a practical problem, the voltage X in a certain electrical system might be a random variable with a continuous distribution that can be approximately represented by the p.d.f.

$$f(x) = \begin{cases} 0 & \text{for } x \le 0, \\ \dfrac{1}{(1+x)^2} & \text{for } x > 0. \end{cases} \tag{3.2.8}$$

It can be verified that the properties (3.2.4) and (3.2.5) required of all p.d.f.'s are satisfied by $f(x)$.

Even though the voltage X may actually be bounded in the real situation, the p.d.f. (3.2.8) may provide a good approximation for the distribution of X over its full range of values. For example, suppose that it is known that the maximum possible value of X is 1000, in which case $\Pr(X > 1000) = 0$. When the p.d.f. (3.2.8) is used, we compute $\Pr(X > 1000) = 0.001$. If (3.2.8) adequately represents the variability of X over the interval (0, 1000), then it may be more convenient to use the p.d.f. (3.2.8) than a p.d.f. that is similar to (3.2.8) for $x \le 1000$, except for a new normalizing

constant, and is 0 for $x > 1000$. This can be especially true if we do not know for sure that the maximum voltage is only 1000. ◄

Example 3.2.6 Unbounded p.d.f.'s. Since a value of a p.d.f. is a probability density, rather than a probability, such a value can be larger than 1. In fact, the values of the following p.d.f. are unbounded in the neighborhood of $x = 0$:

$$f(x) = \begin{cases} \frac{2}{3}x^{-1/3} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.2.9}$$

It can be verified that even though the p.d.f. (3.2.9) is unbounded, it satisfies the properties (3.2.4) and (3.2.5) required of a p.d.f. ◄

Mixed Distributions

Most distributions that are encountered in practical problems are either discrete or continuous. We shall show, however, that it may sometimes be necessary to consider a distribution that is a mixture of a discrete distribution and a continuous distribution.

Example 3.2.7 Truncated Voltage. Suppose that in the electrical system considered in Example 3.2.5, the voltage X is to be measured by a voltmeter that will record the actual value of X if $X \le 3$ but will simply record the value 3 if $X > 3$. If we let Y denote the value recorded by the voltmeter, then the distribution of Y can be derived as follows.

First, $\Pr(Y = 3) = \Pr(X \ge 3) = 1/4$. Since the single value $Y = 3$ has probability 1/4, it follows that $\Pr(0 < Y < 3) = 3/4$. Furthermore, since $Y = X$ for $0 < X < 3$, this probability 3/4 for Y is distributed over the interval (0, 3) according to the same p.d.f. (3.2.8) as that of X over the same interval. Thus, the distribution of Y is specified by the combination of a p.d.f. over the interval (0, 3) and a positive probability at the point $Y = 3$. ◄

Summary

A continuous distribution is characterized by its probability density function (p.d.f.). A nonnegative function f is the p.d.f. of the distribution of X if, for every interval $[a, b]$, $\Pr(a \le X \le b) = \int_a^b f(x)\,dx$. Continuous random variables satisfy $\Pr(X = x) = 0$ for every value x. If the p.d.f. of a distribution is constant on an interval $[a, b]$ and is 0 off the interval, we say that the distribution is uniform on the interval $[a, b]$.

Exercises

1. Let X be a random variable with the p.d.f. specified in Example 3.2.6. Compute $\Pr(X \le 8/27)$.

2. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{4}{3}(1 - x^3) & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Sketch this p.d.f. and determine the values of the following probabilities: **a.** $\Pr\left(X < \frac{1}{2}\right)$ **b.** $\Pr\left(\frac{1}{4} < X < \frac{3}{4}\right)$ **c.** $\Pr\left(X > \frac{1}{3}\right)$.

3. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{1}{36}(9 - x^2) & \text{for } -3 \le x \le 3, \\ 0 & \text{otherwise.} \end{cases}$$

Sketch this p.d.f. and determine the values of the following probabilities: **a.** $\Pr(X < 0)$ **b.** $\Pr(-1 \le X \le 1)$ **c.** $\Pr(X > 2)$.

4. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} cx^2 & \text{for } 1 \le x \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

a. Find the value of the constant c and sketch the p.d.f.

b. Find the value of $\Pr(X > 3/2)$.

5. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{1}{8}x & \text{for } 0 \le x \le 4, \\ 0 & \text{otherwise.} \end{cases}$$

a. Find the value of t such that $\Pr(X \le t) = 1/4$.

b. Find the value of t such that $\Pr(X \ge t) = 1/2$.

6. Let X be a random variable for which the p.d.f. is as given in Exercise 5. After the value of X has been observed, let Y be the integer closest to X. Find the p.f. of the random variable Y.

7. Suppose that a random variable X has the uniform distribution on the interval $[-2, 8]$. Find the p.d.f. of X and the value of $\Pr(0 < X < 7)$.

8. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} ce^{-2x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

a. Find the value of the constant c and sketch the p.d.f.

b. Find the value of $\Pr(1 < X < 2)$.

9. Show that there does not exist any number c such that the following function $f(x)$ would be a p.d.f.:

$$f(x) = \begin{cases} \frac{c}{1+x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

10. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{c}{(1-x)^{1/2}} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

a. Find the value of the constant c and sketch the p.d.f.

b. Find the value of $\Pr(X \le 1/2)$.

11. Show that there does not exist any number c such that the following function $f(x)$ would be a p.d.f.:

$$f(x) = \begin{cases} \frac{c}{x} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

12. In Example 3.1.3 on page 94, determine the distribution of the random variable Y, the electricity demand. Also, find $\Pr(Y < 50)$.

13. An ice cream seller takes 20 gallons of ice cream in her truck each day. Let X stand for the number of gallons that she sells. The probability is 0.1 that $X = 20$. If she doesn't sell all 20 gallons, the distribution of X follows a continuous distribution with a p.d.f. of the form

$$f(x) = \begin{cases} cx & \text{for } 0 < x < 20, \\ 0 & \text{otherwise,} \end{cases}$$

where c is a constant that makes $\Pr(X < 20) = 0.9$. Find the constant c so that $\Pr(X < 20) = 0.9$ as described above.

3.3 The Cumulative Distribution Function

Although a discrete distribution is characterized by its p.f. and a continuous distribution is characterized by its p.d.f., every distribution has a common characterization through its (cumulative) distribution function (c.d.f.). The inverse of the c.d.f. is called the quantile function, and it is useful for indicating where the probability is located in a distribution.

Example 3.3.1

Voltage. Consider again the voltage X from Example 3.2.5. The distribution of X is characterized by the p.d.f. in Eq. (3.2.8). An alternative characterization that is more directly related to probabilities associated with X is obtained from the following function:

$$F(x) = \Pr(X \le x) = \int_{-\infty}^{x} f(y)dy = \begin{cases} 0 & \text{for } x \le 0, \\ \int_0^x \frac{dy}{(1+y)^2} & \text{for } x > 0, \end{cases}$$
$$= \begin{cases} 0 & \text{for } x \le 0, \\ 1 - \frac{1}{1+x} & \text{for } x > 0. \end{cases} \tag{3.3.1}$$

So, for example, $\Pr(X \le 3) = F(3) = 3/4$. ◀

Definition and Basic Properties

Definition 3.3.1 (Cumulative) Distribution Function. The *distribution function* or *cumulative distribution function* (abbreviated *c.d.f.*) F of a random variable X is the function

$$F(x) = \Pr(X \le x) \quad \text{for } -\infty < x < \infty. \tag{3.3.2}$$

It should be emphasized that the cumulative distribution function is defined as above for every random variable X, regardless of whether the distribution of X is discrete, continuous, or mixed. For the continuous random variable in Example 3.3.1, the c.d.f. was calculated in Eq. (3.3.1). Here is a discrete example:

Example 3.3.2 Bernoulli c.d.f. Let X have the Bernoulli distribution with parameter p defined in Definition 3.1.5. Then $\Pr(X = 0) = 1 - p$ and $\Pr(X = 1) = p$. Let F be the c.d.f. of X. It is easy to see that $F(x) = 0$ for $x < 0$ because $X \ge 0$ for sure. Similarly, $F(x) = 1$ for $x \ge 1$ because $X \le 1$ for sure. For $0 \le x < 1$, $\Pr(X \le x) = \Pr(X = 0) = 1 - p$ because 0 is the only possible value of X that is in the interval $(-\infty, x]$. In summary,

$$F(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 - p & \text{for } 0 \le x < 1, \\ 1 & \text{for } x \ge 1. \end{cases}$$ ◀

We shall soon see (Theorem 3.3.2) that the c.d.f. allows calculation of all interval probabilities; hence, it characterizes the distribution of a random variable. It follows from Eq. (3.3.2) that the c.d.f. of each random variable X is a function F defined on the real line. The value of F at every point x must be a number $F(x)$ in the interval [0, 1] because $F(x)$ is the probability of the event $\{X \le x\}$. Furthermore, it follows from Eq. (3.3.2) that the c.d.f. of every random variable X must have the following three properties.

Property 3.3.1 Nondecreasing. *The function $F(x)$ is nondecreasing as x increases; that is, if $x_1 < x_2$, then $F(x_1) \le F(x_2)$.*

Proof If $x_1 < x_2$, then the event $\{X \le x_1\}$ is a subset of the event $\{X \le x_2\}$. Hence, $\Pr\{X \le x_1\} \le \Pr\{X \le x_2\}$ according to Theorem 1.5.4. ■

An example of a c.d.f. is sketched in Fig. 3.6. It is shown in that figure that $0 \le F(x) \le 1$ over the entire real line. Also, $F(x)$ is always nondecreasing as x increases, although $F(x)$ is constant over the interval $x_1 \le x \le x_2$ and for $x \ge x_4$.

Property 3.3.2 Limits at $\pm\infty$. $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$.

Proof As in the proof of Property 3.3.1, note that $\{X \le x_1\} \subset \{X \le x_2\}$ whenever $x_1 < x_2$. The fact that $\Pr(X \le x)$ approaches 0 as $x \to -\infty$ now follows from Exercise 13 in

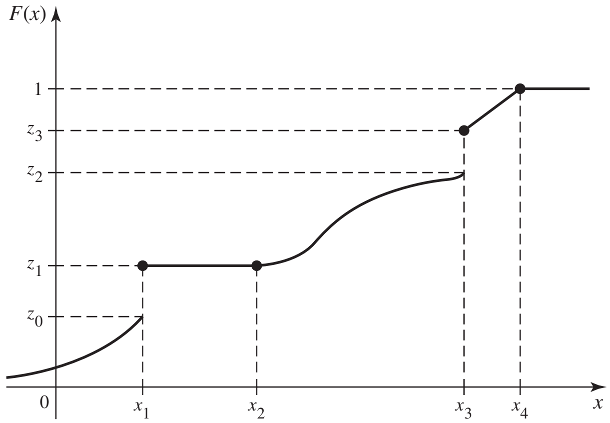


Figure 3.6 An example of a c.d.f.

Section 1.10. Similarly, the fact that $\Pr(X \le x)$ approaches 1 as $x \to \infty$ follows from Exercise 12 in Sec. 1.10. ■

The limiting values specified in Property 3.3.2 are indicated in Fig. 3.6. In this figure, the value of $F(x)$ actually becomes 1 at $x = x_4$ and then remains 1 for $x > x_4$. Hence, it may be concluded that $\Pr(X \le x_4) = 1$ and $\Pr(X > x_4) = 0$. On the other hand, according to the sketch in Fig. 3.6, the value of $F(x)$ approaches 0 as $x \to -\infty$, but does not actually become 0 at any finite point x. Therefore, for every finite value of x, no matter how small, $\Pr(X \le x) > 0$.

A c.d.f. need not be continuous. In fact, the value of $F(x)$ may jump at any finite or countable number of points. In Fig. 3.6, for instance, such jumps or points of discontinuity occur where $x = x_1$ and $x = x_3$. For each fixed value x, we shall let $F(x^-)$ denote the limit of the values of $F(y)$ as y approaches x from the left, that is, as y approaches x through values smaller than x. In symbols,

$$F(x^-) = \lim_{\substack{y \to x \\ y < x}} F(y).$$

Similarly, we shall define $F(x^+)$ as the limit of the values of $F(y)$ as y approaches x from the right. Thus,

$$F(x^+) = \lim_{\substack{y \to x \\ y > x}} F(y).$$

If the c.d.f. is continuous at a given point x, then $F(x^-) = F(x^+) = F(x)$ at that point.

Property 3.3.3 Continuity from the Right. *A c.d.f. is always continuous from the right; that is,* $F(x) = F(x^+)$ *at every point* x.

Proof Let $y_1 > y_2 > \cdots$ be a sequence of numbers that are decreasing such that $\lim_{n\to\infty} y_n = x$. Then the event $\{X \le x\}$ is the intersection of all the events $\{X \le y_n\}$ for $n = 1, 2, \ldots$. Hence, by Exercise 13 of Sec. 1.10,

$$F(x) = \Pr(X \le x) = \lim_{n\to\infty} \Pr(X \le y_n) = F(x^+).$$ ■

It follows from Property 3.3.3 that at every point x at which a jump occurs,

$$F(x^+) = F(x) \text{ and } F(x^-) < F(x).$$

In Fig. 3.6 this property is illustrated by the fact that, at the points of discontinuity $x = x_1$ and $x = x_3$, the value of $F(x_1)$ is taken as z_1 and the value of $F(x_3)$ is taken as z_3.

Determining Probabilities from the Distribution Function

Example 3.3.3 **Voltage.** In Example 3.3.1, suppose that we want to know the probability that X lies in the interval [2, 4]. That is, we want $\Pr(2 \le X \le 4)$. The c.d.f. allows us to compute $\Pr(X \le 4)$ and $\Pr(X \le 2)$. These are related to the probability that we want as follows: Let $A = \{2 < X \le 4\}$, $B = \{X \le 2\}$, and $C = \{X \le 4\}$. Because X has a continuous distribution, $\Pr(A)$ is the same as the probability that we desire. We see that $A \cup B = C$, and it is clear that A and B are disjoint. Hence, $\Pr(A) + \Pr(B) = \Pr(C)$. It follows that

$$\Pr(A) = \Pr(C) - \Pr(B) = F(4) - F(2) = \frac{4}{5} - \frac{3}{4} = \frac{1}{20}.$$ ◀

The type of reasoning used in Example 3.3.3 can be extended to find the probability that an arbitrary random variable X will lie in any specified interval of the real line from the c.d.f. We shall derive this probability for four different types of intervals.

Theorem 3.3.1 For every value x,

$$\Pr(X > x) = 1 - F(x). \tag{3.3.3}$$

Proof The events $\{X > x\}$ and $\{X \le x\}$ are disjoint, and their union is the whole sample space S whose probability is 1. Hence, $\Pr(X > x) + \Pr(X \le x) = 1$. Now, Eq. (3.3.3) follows from Eq. (3.3.2). ■

Theorem 3.3.2 For all values x_1 and x_2 such that $x_1 < x_2$,

$$\Pr(x_1 < X \le x_2) = F(x_2) - F(x_1). \tag{3.3.4}$$

Proof Let $A = \{x_1 < X \le x_2\}$, $B = \{X \le x_1\}$, and $C = \{X \le x_2\}$. As in Example 3.3.3, A and B are disjoint, and their union is C, so

$$\Pr(x_1 < X \le x_2) + \Pr(X \le x_1) = \Pr(X \le x_2).$$

Subtracting $\Pr(X \le x_1)$ from both sides of this equation and applying Eq. (3.3.2) yields Eq. (3.3.4). ■

For example, if the c.d.f. of X is as sketched in Fig. 3.6, then it follows from Theorems 3.3.1 and 3.3.2 that $\Pr(X > x_2) = 1 - z_1$ and $\Pr(x_2 < X \le x_3) = z_3 - z_1$. Also, since $F(x)$ is constant over the interval $x_1 \le x \le x_2$, then $\Pr(x_1 < X \le x_2) = 0$.

It is important to distinguish carefully between the strict inequalities and the weak inequalities that appear in all of the preceding relations and also in the next theorem. If there is a jump in $F(x)$ at a given value x, then the values of $\Pr(X \le x)$ and $\Pr(X < x)$ will be different.

Theorem 3.3.3 For each value x,

$$\Pr(X < x) = F(x^-). \tag{3.3.5}$$

Proof Let $y_1 < y_2 < \cdots$ be an increasing sequence of numbers such that $\lim_{n\to\infty} y_n = x$. Then it can be shown that

$$\{X < x\} = \bigcup_{n=1}^{\infty} \{X \le y_n\}.$$

Therefore, it follows from Exercise 12 of Sec. 1.10 that

$$\begin{aligned}\Pr(X < x) &= \lim_{n\to\infty} \Pr(X \le y_n) \\ &= \lim_{n\to\infty} F(y_n) = F(x^-).\end{aligned}$$ ■

For example, for the c.d.f. sketched in Fig. 3.6, $\Pr(X < x_3) = z_2$ and $\Pr(X < x_4) = 1$.

Finally, we shall show that for every value x, $\Pr(X = x)$ is equal to the amount of the jump that occurs in F at the point x. If F is continuous at the point x, that is, if there is no jump in F at x, then $\Pr(X = x) = 0$.

Theorem 3.3.4 For every value x,

$$\Pr(X = x) = F(x) - F(x^-). \tag{3.3.6}$$

Proof It is always true that $\Pr(X = x) = \Pr(X \le x) - \Pr(X < x)$. The relation (3.3.6) follows from the fact that $\Pr(X \le x) = F(x)$ at every point and from Theorem 3.3.3. ■

In Fig. 3.6, for example, $\Pr(X = x_1) = z_1 - z_0$, $\Pr(X = x_3) = z_3 - z_2$, and the probability of every other individual value of X is 0.

The c.d.f. of a Discrete Distribution

From the definition and properties of a c.d.f. $F(x)$, it follows that if $a < b$ and if $\Pr(a < X < b) = 0$, then $F(x)$ will be constant and horizontal over the interval $a < x < b$. Furthermore, as we have just seen, at every point x such that $\Pr(X = x) > 0$, the c.d.f. will jump by the amount $\Pr(X = x)$.

Suppose that X has a discrete distribution with the p.f. $f(x)$. Together, the properties of a c.d.f. imply that $F(x)$ must have the following form: $F(x)$ will have a jump of magnitude $f(x_i)$ at each possible value x_i of X, and $F(x)$ will be constant between every pair of successive jumps. The distribution of a discrete random variable X can be represented equally well by either the p.f. or the c.d.f. of X.

The c.d.f. of a Continuous Distribution

Theorem 3.3.5 Let X have a continuous distribution, and let $f(x)$ and $F(x)$ denote its p.d.f. and the c.d.f., respectively. Then F is continuous at every x,

$$F(x) = \int_{-\infty}^{x} f(t)\, dt, \tag{3.3.7}$$

and

$$\frac{dF(x)}{dx} = f(x), \tag{3.3.8}$$

at all x such that f is continuous.

Proof Since the probability of each individual point x is 0, the c.d.f. $F(x)$ will have no jumps. Hence, $F(x)$ will be a continuous function over the entire real line.

By definition, $F(x) = \Pr(X \le x)$. Since f is the p.d.f. of X, we have from the definition of p.d.f. that $\Pr(X \le x)$ is the right-hand side of Eq. (3.3.7).

It follows from Eq. (3.3.7) and the relation between integrals and derivatives (the fundamental theorem of calculus) that, for every x at which f is continuous, Eq. (3.3.8) holds. ■

Thus, the c.d.f. of a continuous random variable X can be obtained from the p.d.f. and vice versa. Eq. (3.3.7) is how we found the c.d.f. in Example 3.3.1. Notice that the derivative of the F in Example 3.3.1 is

$$F'(x) = \begin{cases} 0 & \text{for } x < 0, \\ \dfrac{1}{(1+x)^2} & \text{for } x > 0, \end{cases}$$

and F' does not exist at $x = 0$. This verifies Eq (3.3.8) for Example 3.3.1. Here, we have used the popular shorthand notation $F'(x)$ for the derivative of F at the point x.

Example 3.3.4 Calculating a p.d.f. from a c.d.f. Let the c.d.f. of a random variable be

$$F(x) = \begin{cases} 0 & \text{for } x < 0, \\ x^{2/3} & \text{for } 0 \le x \le 1, \\ 1 & \text{for } x > 1. \end{cases}$$

This function clearly satisfies the three properties required of every c.d.f., as given earlier in this section. Furthermore, since this c.d.f. is continuous over the entire real line and is differentiable at every point except $x = 0$ and $x = 1$, the distribution of X is continuous. Therefore, the p.d.f. of X can be found at every point other than $x = 0$ and $x = 1$ by the relation (3.3.8). The value of $f(x)$ at the points $x = 0$ and $x = 1$ can be assigned arbitrarily. When the derivative $F'(x)$ is calculated, it is found that $f(x)$ is as given by Eq. (3.2.9) in Example 3.2.6. Conversely, if the p.d.f. of X is given by Eq. (3.2.9), then by using Eq. (3.3.7) it is found that $F(x)$ is as given in this example. ◄

The Quantile Function

Example 3.3.5 Fair Bets. Suppose that X is the amount of rain that will fall tomorrow, and X has c.d.f. F. Suppose that we want to place an even-money bet on X as follows: If $X \le x_0$, we win one dollar and if $X > x_0$ we lose one dollar. In order to make this bet fair, we need $\Pr(X \le x_0) = \Pr(X > x_0) = 1/2$. We could search through all of the real numbers x trying to find one such that $F(x) = 1/2$, and then we would let x_0 equal the value we found. If F is a one-to-one function, then F has an inverse F^{-1} and $x_0 = F^{-1}(1/2)$. ◄

The value x_0 that we seek in Example 3.3.5 is called the 0.5 *quantile* of X or the 50th *percentile* of X because 50% of the distribution of X is at or below x_0.

Definition 3.3.2 Quantiles/Percentiles. Let X be a random variable with c.d.f. F. For each p strictly between 0 and 1, define $F^{-1}(p)$ to be the smallest value x such that $F(x) \ge p$. Then $F^{-1}(p)$ is called the p *quantile* of X or the $100p$ *percentile* of X. The function F^{-1} defined here on the open interval (0, 1) is called the *quantile function* of X.

Example 3.3.6

Standardized Test Scores. Many universities in the United States rely on standardized test scores as part of their admissions process. Thousands of people take these tests each time that they are offered. Each examinee's score is compared to the collection of scores of all examinees to see where it fits in the overall ranking. For example, if 83% of all test scores are at or below your score, your test report will say that you scored at the 83rd percentile. ◄

The notation $F^{-1}(p)$ in Definition 3.3.2 deserves some justification. Suppose first that the c.d.f. F of X is continuous and one-to-one over the whole set of possible values of X. Then the inverse F^{-1} of F exists, and for each $0 < p < 1$, there is one and only one x such that $F(x) = p$. That x is $F^{-1}(p)$. Definition 3.3.2 extends the concept of inverse function to nondecreasing functions (such as c.d.f.'s) that may be neither one-to-one nor continuous.

Quantiles of Continuous Distributions When the c.d.f. of a random variable X is continuous and one-to-one over the whole set of possible values of X, the inverse F^{-1} of F exists and equals the quantile function of X.

Example 3.3.7

Value at Risk. The manager of an investment portfolio is interested in how much money the portfolio might lose over a fixed time horizon. Let X be the change in value of the given portfolio over a period of one month. Suppose that X has the p.d.f. in Fig. 3.7. The manager computes a quantity known in the world of risk management as *Value at Risk* (denoted by VaR). To be specific, let $Y = -X$ stand for the loss incurred by the portfolio over the one month. The manager wants to have a level of confidence about how large Y might be. In this example, the manager specifies a probability level, such as 0.99 and then finds y_0, the 0.99 quantile of Y. The manager is now 99% sure that $Y \le y_0$, and y_0 is called the VaR. If X has a continuous distribution, then it is easy to see that y_0 is closely related to the 0.01 quantile of the distribution of X. The 0.01 quantile x_0 has the property that $\Pr(X < x_0) = 0.01$. But $\Pr(X < x_0) = \Pr(Y > -x_0) = 1 - \Pr(Y \le -x_0)$. Hence, $-x_0$ is a 0.99 quantile of Y. For the p.d.f. in Fig. 3.7, we see that $x_0 = -4.14$, as the shaded region indicates. Then $y_0 = 4.14$ is VaR for one month at probability level 0.99. ◄

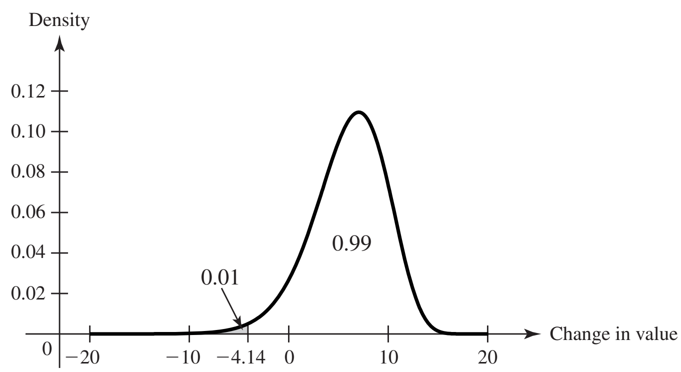


Figure 3.7 The p.d.f. of the change in value of a portfolio with lower 1% indicated.

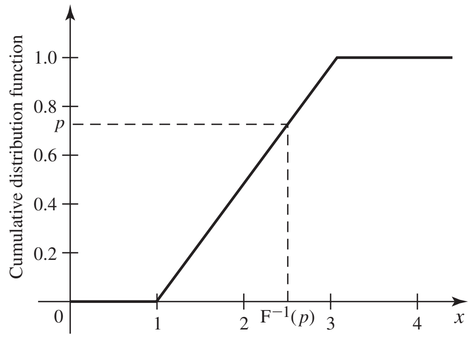


Figure 3.8 The c.d.f. of a uniform distribution indicating how to solve for a quantile.

Example 3.3.8

Uniform Distribution on an Interval. Let X have the uniform distribution on the interval $[a, b]$. The c.d.f. of X is

$$F(x) = \Pr(X \le x) = \begin{cases} 0 & \text{if } x \le a, \\ \displaystyle\int_a^x \frac{1}{b-a} du & \text{if } a < x \le b, \\ 1 & \text{if } x > b. \end{cases}$$

The integral above equals $(x-a)/(b-a)$. So, $F(x) = (x-a)/(b-a)$ for all $a < x < b$, which is a strictly increasing function over the entire interval of possible values of X. The inverse of this function is the quantile function of X, which we obtain by setting $F(x)$ equal to p and solving for x:

$$\frac{x-a}{b-a} = p,$$

$$x - a = p(b-a),$$

$$x = a + p(b-a) = pb + (1-p)a.$$

Figure 3.8 illustrates how the calculation of a quantile relates to the c.d.f.

The quantile function of X is $F^{-1}(p) = pb + (1-p)a$ for $0 < p < 1$. In particular, $F^{-1}(1/2) = (b+a)/2$. ◀

Note: Quantiles, Like c.d.f.'s, Depend on the Distribution Only. Any two random variables with the same distribution have the same quantile function. When we refer to a quantile of X, we mean a quantile of the distribution of X.

Quantiles of Discrete Distributions It is convenient to be able to calculate quantiles for discrete distributions as well. The quantile function of Definition 3.3.2 exists for all distributions whether discrete, continuous, or otherwise. For example, in Fig. 3.6, let $z_0 \le p \le z_1$. Then the smallest x such that $F(x) \ge p$ is x_1. For every value of $x < x_1$, we have $F(x) < z_0 \le p$ and $F(x_1) = z_1$. Notice that $F(x) = z_1$ for all x between x_1 and x_2, but since x_1 is the smallest of all those numbers, x_1 is the p quantile. Because distribution functions are continuous from the right, the smallest x such that $F(x) \ge p$ exists for all $0 < p < 1$. For $p = 1$, there is no guarantee that such an x will exist. For example, in Fig. 3.6, $F(x_4) = 1$, but in Example 3.3.1, $F(x) < 1$ for all x. For $p = 0$, there is never a smallest x such that $F(x) = 0$ because $\lim_{x\to-\infty} F(x) = 0$. That is, if $F(x_0) = 0$, then $F(x) = 0$ for all $x < x_0$. For these reasons, we never talk about the 0 or 1 quantiles.

Table 3.1 Quantile function for Example 3.3.9

p	$F^{-1}(p)$
(0, 0.1681]	0
(0.1681, 0.5283]	1
(0.5283, 0.8370]	2
(0.8370, 0.9693]	3
(0.9693, 0.9977]	4
(0.9977, 1)	5

Example 3.3.9

Quantiles of a Binomial Distribution. Let X have the binomial distribution with parameters 5 and 0.3. The binomial table in the back of the book has the p.f. f of X, which we reproduce here together with the c.d.f. F:

x	0	1	2	3	4	5
$f(x)$	0.1681	0.3602	0.3087	0.1323	0.0284	0.0024
$F(x)$	0.1681	0.5283	0.8370	0.9693	0.9977	1

(A little rounding error occurred in the p.f.) So, for example, the 0.5 quantile of this distribution is 1, which is also the 0.25 quantile and the 0.20 quantile. The entire quantile function is in Table 3.1. So, the 90th percentile is 3, which is also the 95th percentile, etc. ◄

Certain quantiles have special names.

Definition 3.3.3

Median/Quartiles. The 1/2 quantile or the 50th percentile of a distribution is called its *median*. The 1/4 quantile or 25th percentile is the *lower quartile*. The 3/4 quantile or 75th percentile is called the *upper quartile*.

Note: The Median Is Special. The median of a distribution is one of several special features that people like to use when sumarizing the distribution of a random variable. We shall discuss summaries of distributions in more detail in Chapter 4. Because the median is such a popular summary, we need to note that there are several different but similar "definitions" of median. Recall that the 1/2 quantile is the *smallest* number x such that $F(x) \geq 1/2$. For some distributions, usually discrete distributions, there will be an interval of numbers $[x_1, x_2)$ such that for all $x \in [x_1, x_2)$, $F(x) = 1/2$. In such cases, it is common to refer to all such x (including x_2) as medians of the distribution. (See Definition 4.5.1.) Another popular convention is to call $(x_1 + x_2)/2$ the median. This last is probably the most common convention. The readers should be aware that, whenever they encounter a median, it might be any one of the things that we just discussed. Fortunately, they all mean nearly the same thing, namely that the number divides the distribution in half as closely as is possible.

Example 3.3.10 **Uniform Distribution on Integers.** Let X have the uniform distribution on the integers 1, 2, 3, 4. (See Definition 3.1.6.) The c.d.f. of X is

$$F(x) = \begin{cases} 0 & \text{if } x < 1, \\ 1/4 & \text{if } 1 \le x < 2, \\ 1/2 & \text{if } 2 \le x < 3, \\ 3/4 & \text{if } 3 \le x < 4, \\ 1 & \text{if } x \ge 4. \end{cases}$$

The 1/2 quantile is 2, but every number in the interval [2, 3] might be called a median. The most popular choice would be 2.5. ◂

One advantage to describing a distribution by the quantile function rather than by the c.d.f. is that quantile functions are easier to display in tabular form for multiple distributions. The reason is that the domain of the quantile function is always the interval (0, 1) no matter what the possible values of X are. Quantiles are also useful for summarizing distributions in terms of where the probability is. For example, if one wishes to say where the middle half of a distribution is, one can say that it lies between the 0.25 quantile and the 0.75 quantile. In Sec. 8.5, we shall see how to use quantiles to help provide estimates of unknown quantities after observing data.

In Exercise 19, you can show how to recover the c.d.f. from the quantile function. Hence, the quantile function is an alternative way to characterize a distribution.

Summary

The c.d.f. F of a random variable X is $F(x) = \Pr(X \le x)$ for all real x. This function is continuous from the right. If we let $F(x^-)$ equal the limit of $F(y)$ as y approaches x from below, then $F(x) - F(x^-) = \Pr(X = x)$. A continuous distribution has a continuous c.d.f. and $F'(x) = f(x)$, the p.d.f. of the distribution, for all x at which F is differentiable. A discrete distribution has a c.d.f. that is constant between the possible values and jumps by $f(x)$ at each possible value x. The quantile function $F^{-1}(p)$ is equal to the smallest x such that $F(x) \ge p$ for $0 < p < 1$.

Exercises

1. Suppose that a random variable X has the Bernoulli distribution with parameter $p = 0.7$. (See Definition 3.1.5.) Sketch the c.d.f. of X.

2. Suppose that a random variable X can take only the values -2, 0, 1, and 4, and that the probabilities of these values are as follows: $\Pr(X = -2) = 0.4$, $\Pr(X = 0) = 0.1$, $\Pr(X = 1) = 0.3$, and $\Pr(X = 4) = 0.2$. Sketch the c.d.f. of X.

3. Suppose that a coin is tossed repeatedly until a head is obtained for the first time, and let X denote the number of tosses that are required. Sketch the c.d.f. of X.

4. Suppose that the c.d.f. F of a random variable X is as sketched in Fig. 3.9. Find each of the following probabilities:

a. $\Pr(X = -1)$ **b.** $\Pr(X < 0)$
c. $\Pr(X \le 0)$ **d.** $\Pr(X = 1)$
e. $\Pr(0 < X \le 3)$ **f.** $\Pr(0 < X < 3)$
g. $\Pr(0 \le X \le 3)$ **h.** $\Pr(1 < X \le 2)$
i. $\Pr(1 \le X \le 2)$ **j.** $\Pr(X > 5)$
k. $\Pr(X \ge 5)$ **l.** $\Pr(3 \le X \le 4)$

5. Suppose that the c.d.f. of a random variable X is as follows:

$$F(x) = \begin{cases} 0 & \text{for } x \le 0, \\ \frac{1}{9}x^2 & \text{for } 0 < x \le 3, \\ 1 & \text{for } x > 3. \end{cases}$$

Find and sketch the p.d.f. of X.

6. Suppose that the c.d.f. of a random variable X is as follows:

$$F(x) = \begin{cases} e^{x-3} & \text{for } x \le 3, \\ 1 & \text{for } x > 3. \end{cases}$$

Find and sketch the p.d.f. of X.

7. Suppose, as in Exercise 7 of Sec. 3.2, that a random variable X has the uniform distribution on the interval $[-2, 8]$. Find and sketch the c.d.f. of X.

8. Suppose that a point in the xy-plane is chosen at random from the interior of a circle for which the equation is $x^2 + y^2 = 1$; and suppose that the probability that the point will belong to each region inside the circle is proportional to the area of that region. Let Z denote a random variable representing the distance from the center of the circle to the point. Find and sketch the c.d.f. of Z.

9. Suppose that X has the uniform distribution on the interval $[0, 5]$ and that the random variable Y is defined by $Y = 0$ if $X \le 1$, $Y = 5$ if $X \ge 3$, and $Y = X$ otherwise. Sketch the c.d.f. of Y.

10. For the c.d.f. in Example 3.3.4, find the quantile function.

11. For the c.d.f. in Exercise 5, find the quantile function.

12. For the c.d.f. in Exercise 6, find the quantile function.

13. Suppose that a broker believes that the change in value X of a particular investment over the next two months has the uniform distribution on the interval $[-12, 24]$. Find the value at risk VaR for two months at probability level 0.95.

14. Find the quartiles and the median of the binomial distribution with parameters $n = 10$ and $p = 0.2$.

15. Suppose that X has the p.d.f.

$$f(x) = \begin{cases} 2x & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find and sketch the c.d.f. or X.

16. Find the quantile function for the distribution in Example 3.3.1.

17. Prove that the quantile function F^{-1} of a general random variable X has the following three properties that are analogous to properties of the c.d.f.:

a. F^{-1} is a nondecreasing function of p for $0 < p < 1$.

b. Let $x_0 = \lim_{\substack{p \to 0 \\ p > 0}} F^{-1}(p)$ and $x_1 = \lim_{\substack{p \to 1 \\ p < 1}} F^{-1}(p)$. Then x_0 equals the greatest lower bound on the set of numbers c such that $\Pr(X \le c) > 0$, and x_1 equals the least upper bound on the set of numbers d such that $\Pr(X \ge d) > 0$.

c. F^{-1} is continuous from the left; that is $F^{-1}(p) = F^{-1}(p^-)$ for all $0 < p < 1$.

18. Let X be a random variable with quantile function F^{-1}. Assume the following three conditions: (i) $F^{-1}(p) = c$ for all p in the interval (p_0, p_1), (ii) either $p_0 = 0$ or $F^{-1}(p_0) < c$, and (iii) either $p_1 = 1$ or $F^{-1}(p) > c$ for $p > p_1$. Prove that $\Pr(X = c) = p_1 - p_0$.

19. Let X be a random variable with c.d.f. F and quantile function F^{-1}. Let x_0 and x_1 be as defined in Exercise 17. (Note that $x_0 = -\infty$ and/or $x_1 = \infty$ are possible.) Prove that for all x in the open interval (x_0, x_1), $F(x)$ is the largest p such that $F^{-1}(p) \le x$.

20. In Exercise 13 of Sec. 3.2, draw a sketch of the c.d.f. F of X and find $F(10)$.

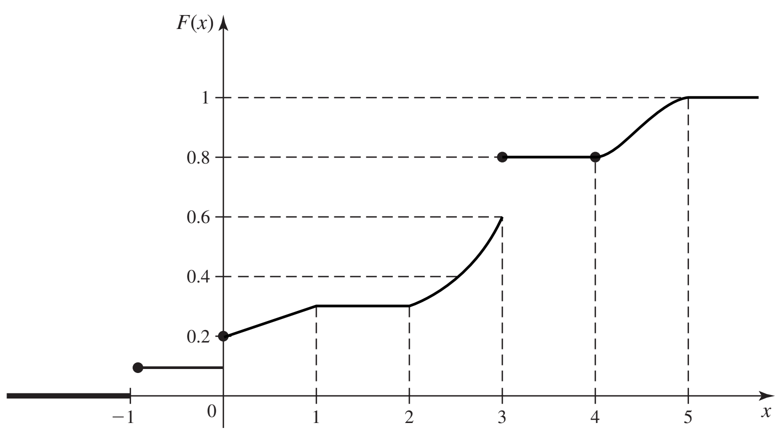


Figure 3.9 The c.d.f. for Exercise 4.

3.4 Bivariate Distributions

We generalize the concept of distribution of a random variable to the joint distribution of two random variables. In doing so, we introduce the joint p.f. for two discrete random variables, the joint p.d.f. for two continuous random variables, and the joint c.d.f. for any two random variables. We also introduce a joint hybrid of p.f. and p.d.f. for the case of one discrete random variable and one continuous random variable.

Example 3.4.1

Demands for Utilities. In Example 3.1.5, we found the distribution of the random variable X that represented the demand for water. But there is another random variable, Y, the demand for electricity, that is also of interest. When discussing two random variables at once, it is often convenient to put them together into an ordered pair, (X, Y). As early as Example 1.5.4 on page 19, we actually calculated some probabilities associated with the pair (X, Y). In that example, we defined two events A and B that we now can express as $A = \{X \geq 115\}$ and $B = \{Y \geq 110\}$. In Example 1.5.4, we computed $\Pr(A \cap B)$ and $\Pr(A \cup B)$. We can express $A \cap B$ and $A \cup B$ as events involving the pair (X, Y). For example, define the set of ordered pairs $C = \{(x, y) : x \geq 115 \text{ and } y \geq 110\}$ so that that the event $\{(X, Y) \in C)\} = A \cap B$. That is, the event that the pair of random variables lies in the set C is the same as the intersection of the two events A and B. In Example 1.5.4, we computed $\Pr(A \cap B) = 0.1198$. So, we can now assert that $\Pr((X, Y) \in C) = 0.1198$. ◀

Definition 3.4.1

Joint/Bivariate Distribution. Let X and Y be random variables. The *joint distribution* or *bivariate distribution* of X and Y is the collection of all probabilities of the form $\Pr[(X, Y) \in C]$ for all sets C of pairs of real numbers such that $\{(X, Y) \in C\}$ is an event.

It is a straightforward consequence of the definition of the joint distribution of X and Y that this joint distribution is itself a probability measure on the set of ordered pairs of real numbers. The set $\{(X, Y) \in C\}$ will be an event for every set C of pairs of real numbers that most readers will be able to imagine.

In this section and the next two sections, we shall discuss convenient ways to characterize and do computations with bivariate distributions. In Sec. 3.7, these considerations will be extended to the joint distribution of an arbitrary finite number of random variables.

Discrete Joint Distributions

Example 3.4.2

Theater Patrons. Suppose that a sample of 10 people is selected at random from a theater with 200 patrons. One random variable of interest might be the number X of people in the sample who are over 60 years of age, and another random variable might be the number Y of people in the sample who live more than 25 miles from the theater. For each ordered pair (x, y) with $x = 0, \dots, 10$ and $y = 0, \dots, 10$, we might wish to compute $\Pr((X, Y) = (x, y))$, the probability that there are x people in the sample who are over 60 years of age and there are y people in the sample who live more than 25 miles away. ◀

Definition 3.4.2

Discrete Joint Distribution. Let X and Y be random variables, and consider the ordered pair (X, Y). If there are only finitely or at most countably many different possible values (x, y) for the pair (X, Y), then we say that X and Y have a *discrete joint distribution*.

The two random variables in Example 3.4.2 have a discrete joint distribution.

Theorem 3.4.1 Suppose that two random variables X and Y each have a discrete distribution. Then X and Y have a discrete joint distribution.

Proof If both X and Y have only finitely many possible values, then there will be only a finite number of different possible values (x, y) for the pair (X, Y). On the other hand, if either X or Y or both can take a countably infinite number of possible values, then there will also be a countably infinite number of possible values for the pair (X, Y). In all of these cases, the pair (X, Y) has a discrete joint distribution. ■

When we define continuous joint distribution shortly, we shall see that the obvious analog of Theorem 3.4.1 is not true.

Definition 3.4.3 Joint Probability Function, p.f. The *joint probability function*, or the *joint p.f.*, of X and Y is defined as the function f such that for every point (x, y) in the xy-plane,

$$f(x, y) = \Pr(X = x \text{ and } Y = y).$$

The following result is easy to prove because there are at most countably many pairs (x, y) that must account for all of the probability a discrete joint distribution.

Theorem 3.4.2 Let X and Y have a discrete joint distribution. If (x, y) is not one of the possible values of the pair (X, Y), then $f(x, y) = 0$. Also,

$$\sum_{\text{All } (x,y)} f(x, y) = 1.$$

Finally, for each set C of ordered pairs,

$$\Pr[(X, Y) \in C] = \sum_{(x,y)\in C} f(x, y). \qquad ■$$

Example 3.4.3 Specifying a Discrete Joint Distribution by a Table of Probabilities. In a certain suburban area, each household reported the number of cars and the number of television sets that they owned. Let X stand for the number of cars owned by a randomly selected household in this area. Let Y stand for the number of television sets owned by that same randomly selected household. In this case, X takes only the values 1, 2, and 3; Y takes only the values 1, 2, 3, and 4; and the joint p.f. f of X and Y is as specified in Table 3.2.

Table 3.2 Joint p.f. $f(x, y)$ for Example 3.4.3

	y			
x	1	2	3	4
1	0.1	0	0.1	0
2	0.3	0	0.1	0.2
3	0	0.2	0	0

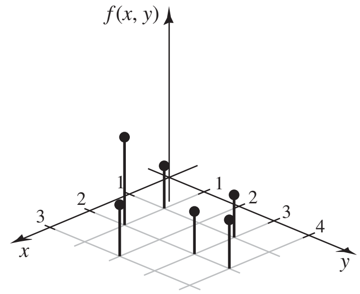


Figure 3.10 The joint p.f. of X and Y in Example 3.4.3.

This joint p.f. is sketched in Fig. 3.10. We shall determine the probability that the randomly selected household owns at least two of both cars and televisions. In symbols, this is $\Pr(X \geq 2 \text{ and } Y \geq 2)$.

By summing $f(x, y)$ over all values of $x \geq 2$ and $y \geq 2$, we obtain the value

$$\begin{aligned}\Pr(X \geq 2 \text{ and } Y \geq 2) &= f(2, 2) + f(2, 3) + f(2, 4) + f(3, 2)\\ &\quad + f(3, 3) + f(3, 4)\\ &= 0.5.\end{aligned}$$

Next, we shall determine the probability that the randomly selected household owns exactly one car, namely $\Pr(X = 1)$. By summing the probabilities in the first row of the table, we obtain the value

$$\Pr(X = 1) = \sum_{y=1}^{4} f(1, y) = 0.2. \quad ◀$$

Continuous Joint Distributions

Example 3.4.4 **Demands for Utilities.** Consider again the joint distribution of X and Y in Example 3.4.1. When we first calculated probabilities for these two random variables back in Example 1.5.4 on page 19 (even before we named them or called them random variables), we assumed that the probability of each subset of the sample space was proportional to the area of the subset. Since the area of the sample space is 29,204, the probability that the pair (X, Y) lies in a region C is the area of C divided by 29,204. We can also write this relation as

$$\Pr((X, Y) \in C\} = \int_C \int \frac{1}{29{,}204}\, dx\, dy, \tag{3.4.1}$$

assuming that the integral exists. ◀

If one looks carefully at Eq. (3.4.1), one will notice the similarity to Eqs. (3.2.2) and (3.2.1). We formalize this connection as follows.

Definition 3.4.4 **Continuous Joint Distribution/Joint p.d.f./Support.** Two random variables X and Y have a *continuous joint distribution* if there exists a nonnegative function f defined over the entire xy-plane such that for every subset C of the plane,

$$\Pr[(X, Y) \in C] = \int_C \int f(x, y)\, dx\, dy,$$

if the integral exists. The function f is called the *joint probability density function* (abbreviated *joint p.d.f.*) of X and Y. The closure of the set $\{(x, y) : f(x, y) > 0\}$ is called the *support of (the distribution of)* (X, Y).

Example 3.4.5

Demands for Utilities. In Example 3.4.4, it is clear from Eq. (3.4.1) that the joint p.d.f. of X and Y is the function

$$f(x, y) = \begin{cases} \dfrac{1}{29{,}204} & \text{for } 4 \le x \le 200 \text{ and } 1 \le y \le 150, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4.2}$$

◀

It is clear from Definition 3.4.4 that the joint p.d.f. of two random variables characterizes their joint distribution. The following result is also straightforward.

Theorem 3.4.3

A joint p.d.f. must satisfy the following two conditions:

$$f(x, y) \ge 0 \quad \text{for } -\infty < x < \infty \text{ and } -\infty < y < \infty,$$

and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1.$$

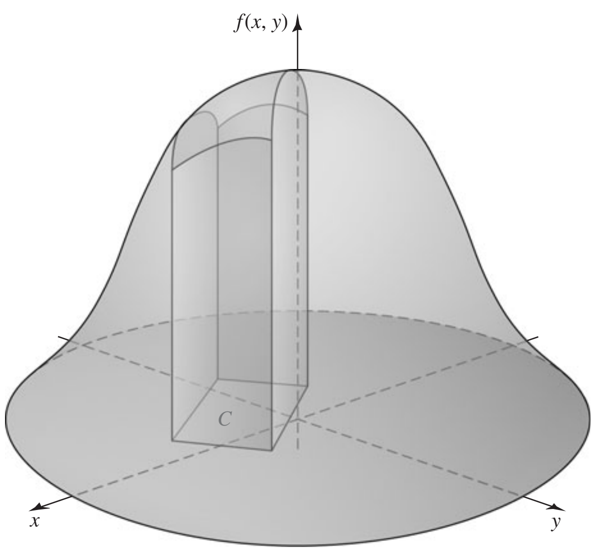
■

Any function that satisfies the two displayed formulas in Theorem 3.4.3 is the joint p.d.f. for some probability distribution.

An example of the graph of a joint p.d.f. is presented in Fig. 3.11. The total volume beneath the surface $z = f(x, y)$ and above the xy-plane must be 1. The probability that the pair (X, Y) will belong to the rectangle C is equal to the volume of the solid figure with base A shown in Fig. 3.11. The top of this solid figure is formed by the surface $z = f(x, y)$.

In Sec. 3.5, we will show that if X and Y have a continuous joint distribution, then X and Y each have a continuous distribution when considered separately. This seems reasonable intutively. However, the converse of this statement is not true, and the following result helps to show why.

Figure 3.11 An example of a joint p.d.f.



Theorem 3.4.4 For every continuous joint distribution on the xy-plane, the following two statements hold:

i. Every individual point, and every infinite sequence of points, in the xy-plane has probability 0.
ii. Let f be a continuous function of one real variable defined on a (possibly unbounded) interval (a, b). The sets $\{(x, y) : y = f(x), a < x < b\}$ and $\{(x, y) : x = f(y), a < y < b\}$ have probability 0.

Proof According to Definition 3.4.4, the probability that a continuous joint distribution assigns to a specified region of the xy-plane can be found by integrating the joint p.d.f. $f(x, y)$ over that region, if the integral exists. If the region is a single point, the integral will be 0. By Axiom 3 of probability, the probability for any countable collection of points must also be 0. The integral of a function of two variables over the graph of a continuous function in the xy-plane is also 0. ■

Example 3.4.6 Not a Continuous Joint Distribution. It follows from (ii) of Theorem 3.4.4 that the probability that (X, Y) will lie on each specified straight line in the plane is 0. If X has a continuous distribution and if $Y = X$, then both X and Y have continuous distributions, but the probability is 1 that (X, Y) lies on the straight line $y = x$. Hence, X and Y cannot have a continuous joint distribution. ◀

Example 3.4.7 Calculating a Normalizing Constant. Suppose that the joint p.d.f. of X and Y is specified as follows:

$$f(x, y) = \begin{cases} cx^2y & \text{for } x^2 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We shall determine the value of the constant c.

The support S of (X, Y) is sketched in Fig. 3.12. Since $f(x, y) = 0$ outside S, it follows that

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy &= \int_S\int f(x, y)\, dx\, dy \\ &= \int_{-1}^{1}\int_{x^2}^{1} cx^2y\, dy\, dx = \frac{4}{21}c. \end{aligned} \tag{3.4.3}$$

Since the value of this integral must be 1, the value of c must be 21/4.

The limits of integration on the last integral in (3.4.3) were determined as follows. We have our choice of whether to integrate x or y as the inner integral, and we chose y. So, we must find, for each x, the interval of y values over which to integrate. From Fig. 3.12, we see that, for each x, y runs from the curve where $y = x^2$ to the line where $y = 1$. The interval of x values for the outer integral is from -1 to 1 according to Fig. 3.12. If we had chosen to integrate x on the inside, then for each y, we see that x runs from $-\sqrt{y}$ to $\sqrt{y}$, while y runs from 0 to 1. The final answer would have been the same. ◀

Example 3.4.8 Calculating Probabilities from a Joint p.d.f. For the joint distribution in Example 3.4.7, we shall now determine the value of $\Pr(X \ge Y)$.

The subset S_0 of S where $x \ge y$ is sketched in Fig. 3.13. Hence,

$$\Pr(X \ge Y) = \int_{S_0}\int f(x, y)\, dx\, dy = \int_0^1\int_{x^2}^{x} \frac{21}{4}x^2y\, dy\, dx = \frac{3}{20}.$$ ◀

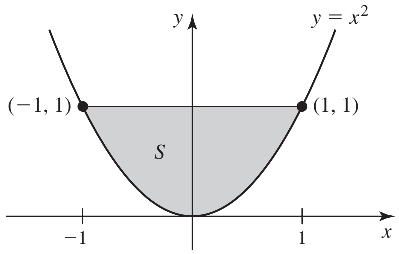


Figure 3.12 The support S of (X, Y) in Example 3.4.8.

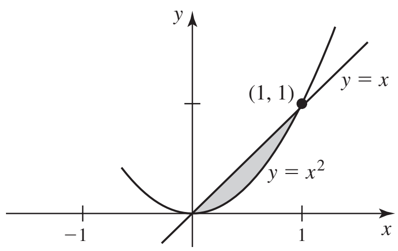


Figure 3.13 The subset S_0 of the support S where $x \geq y$ in Example 3.4.8.

Example 3.4.9

Determining a Joint p.d.f. by Geometric Methods. Suppose that a point (X, Y) is selected at random from inside the circle $x^2 + y^2 \leq 9$. We shall determine the joint p.d.f. of X and Y.

The support of (X, Y) is the set S of points on and inside the circle $x^2 + y^2 \leq 9$. The statement that the point (X, Y) is selected at random from inside the circle is interpreted to mean that the joint p.d.f. of X and Y is constant over S and is 0 outside S. Thus,

$$f(x, y) = \begin{cases} c & \text{for } (x, y) \in S, \\ 0 & \text{otherwise.} \end{cases}$$

We must have

$$\int_S \int f(x, y)\, dx\, dy = c \times (\text{area of } S) = 1.$$

Since the area of the circle S is 9π, the value of the constant c must be $1/(9\pi)$. ◀

Mixed Bivariate Distributions

Example 3.4.10

A Clinical Trial. Consider a clinical trial (such as the one described in Example 2.1.12) in which each patient with depression receives a treatment and is followed to see whether they have a relapse into depression. Let X be the indicator of whether or not the first patient is a "success" (no relapse). That is $X = 1$ if the patient does not relapse and $X = 0$ if the patient relapses. Also, let P be the proportion of patients who have no replapse among all patients who might receive the treatment. It is clear that X must have a discrete distribution, but it might be sensible to think of P as a continuous random variable taking its value anywhere in the interval [0, 1]. Even though X and P can have neither a joint discrete distribution nor a joint continuous distribution, we can still be interested in the joint distribution of X and P. ◀

Prior to Example 3.4.10, we have discussed bivariate distributions that were either discrete or continuous. Occasionally, one must consider a mixed bivariate distribution in which one of the random variables is discrete and the other is continuous. We shall use a function $f(x, y)$ to characterize such a joint distribution in much the same way that we use a joint p.f. to characterize a discrete joint distribution or a joint p.d.f. to characterize a continuous joint distribution.

Definition 3.4.5 Joint p.f./p.d.f. Let X and Y be random variables such that X is discrete and Y is continuous. Suppose that there is a function $f(x, y)$ defined on the xy-plane such that, for every pair A and B of subsets of the real numbers,

$$\Pr(X \in A \text{ and } Y \in B) = \int_B \sum_{x \in A} f(x, y)\,dy, \tag{3.4.4}$$

if the integral exists. Then the function f is called the *joint p.f./p.d.f.* of X and Y.

Clearly, Definition 3.4.5 can be modified in an obvious way if Y is discrete and X is continuous. Every joint p.f./p.d.f. must satisfy two conditions. If X is the discrete random variable with possible values $x_1, x_2, \ldots$ and Y is the continuous random variable, then $f(x, y) \ge 0$ for all x, y and

$$\int_{-\infty}^{\infty} \sum_{i=1}^{\infty} f(x_i, y)\,dy = 1. \tag{3.4.5}$$

Because f is nonnegative, the sum and integral in Eqs. (3.4.4) and (3.4.5) can be done in whichever order is more convenient.

Note: Probabilities of More General Sets. For a general set C of pairs of real numbers, we can compute $\Pr((X, Y) \in C)$ using the joint p.f./p.d.f. of X and Y. For each x, let $C_x = \{y : (x, y) \in C\}$. Then

$$\Pr((X, Y) \in C) = \sum_{\text{All } x} \int_{C_x} f(x, y)\,dy,$$

if all of the integrals exist. Alternatively, for each y, define $C^y = \{x : (x, y) \in C\}$, and then

$$\Pr((X, Y) \in C) = \int_{-\infty}^{\infty} \left[\sum_{x \in C^y} f(x, y) \right] dy,$$

if the integral exists.

Example 3.4.11 A Joint p.f./p.d.f. Suppose that the joint p.f./p.d.f. of X and Y is

$$f(x, y) = \frac{xy^{x-1}}{3}, \quad \text{for } x = 1, 2, 3 \text{ and } 0 < y < 1.$$

We should check to make sure that this function satisfies (3.4.5). It is easier to integrate over the y values first, so we compute

$$\sum_{x=1}^{3} \int_0^1 \frac{xy^{x-1}}{3}\,dy = \sum_{x=1}^{3} \frac{1}{3} = 1.$$

Suppose that we wish to compute the probability that $Y \ge 1/2$ and $X \ge 2$. That is, we want $\Pr(X \in A \text{ and } Y \in B)$ with $A = [2, \infty)$ and $B = [1/2, \infty)$. So, we apply Eq. (3.4.4)

to get the probability

$$\sum_{x=2}^{3}\int_{1/2}^{1}\frac{xy^{x-1}}{3}dy=\sum_{x=2}^{3}\left(\frac{1-(1/2)^x}{3}\right)=0.5417.$$

For illustration, we shall compute the sum and integral in the other order also. For each $y \in [1/2, 1)$, $\sum_{x=2}^{3} f(x, y) = 2y/3 + y^2$. For $y \geq 1/2$, the sum is 0. So, the probability is

$$\int_{1/2}^{1}\left[\frac{2}{3}y+y^2\right]dy=\frac{1}{3}\left[1-\left(\frac{1}{2}\right)^2\right]+\frac{1}{3}\left[1-\left(\frac{1}{2}\right)^3\right]=0.5417.$$ ◂

Example 3.4.12 **A Clinical Trial.** A possible joint p.f./p.d.f. for X and P in Example 3.4.10 is

$$f(x, p) = p^x(1-p)^{1-x}, \quad \text{for } x = 0, 1 \text{ and } 0 < p < 1.$$

Here, X is discrete and P is continuous. The function f is nonnegative, and the reader should be able to demonstrate that it satisfies (3.4.5). Suppose that we wish to compute $\Pr(X \leq 0 \text{ and } P \leq 1/2)$. This can be computed as

$$\int_0^{1/2}(1-p)dp=-\frac{1}{2}[(1-1/2)^2-(1-0)^2]=\frac{3}{8}.$$

Suppose that we also wish to compute $\Pr(X = 1)$. This time, we apply Eq. (3.4.4) with $A = \{1\}$ and $B = (0, 1)$. In this case,

$$\Pr(X=1)=\int_0^1 p\,dp=\frac{1}{2}.$$ ◂

A more complicated type of joint distribution can also arise in a practical problem.

Example 3.4.13 **A Complicated Joint Distribution.** Suppose that X and Y are the times at which two specific components in an electronic system fail. There might be a certain probability p $(0 < p < 1)$ that the two components will fail at the same time and a certain probability $1 - p$ that they will fail at different times. Furthermore, if they fail at the same time, then their common failure time might be distributed according to a certain p.d.f. $f(x)$; if they fail at different times, then these times might be distributed according to a certain joint p.d.f. $g(x, y)$.

The joint distribution of X and Y in this example is not continuous, because there is positive probability p that (X, Y) will lie on the line $x = y$. Nor does the joint distribution have a joint p.f./p.d.f. or any other simple function to describe it. There are ways to deal with such joint distributions, but we shall not discuss them in this text. ◂

Bivariate Cumulative Distribution Functions

The first calculation in Example 3.4.12, namely, $\Pr(X \leq 0 \text{ and } Y \leq 1/2)$, is a generalization of the calculation of a c.d.f. to a bivariate distribution. We formalize the generalization as follows.

Definition 3.4.6 **Joint (Cumulative) Distribution Function/c.d.f.** The *joint distribution function* or *joint cumulative distribution function* (*joint c.d.f.*) of two random variables X and Y is

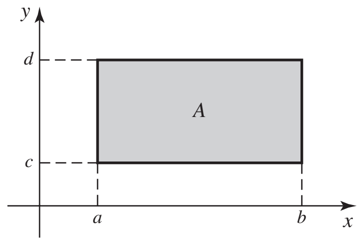


Figure 3.14 The probability of a rectangle.

defined as the function F such that for all values of x and y ($-\infty < x < \infty$ and $-\infty < y < \infty$),

$$F(x, y) = \Pr(X \le x \text{ and } Y \le y).$$

It is clear from Definition 3.4.6 that $F(x, y)$ is monotone increasing in x for each fixed y and is monotone increasing in y for each fixed x.

If the joint c.d.f. of two arbitrary random variables X and Y is F, then the probability that the pair (X, Y) will lie in a specified rectangle in the xy-plane can be found from F as follows: For given numbers $a < b$ and $c < d$,

$$\begin{aligned}
&\Pr(a < X \le b \text{ and } c < Y \le d) \\
&\quad = \Pr(a < X \le b \text{ and } Y \le d) - \Pr(a < X \le b \text{ and } Y \le c) \\
&\quad = [\Pr(X \le b \text{ and } Y \le d) - \Pr(X \le a \text{ and } Y \le d)] \\
&\qquad -[\Pr(X \le b \text{ and } Y \le c) - \Pr(X \le a \text{ and } Y \le c)] \\
&\quad = F(b, d) - F(a, d) - F(b, c) + F(a, c).
\end{aligned} \tag{3.4.6}$$

Hence, the probability of the rectangle C sketched in Fig. 3.14 is given by the combination of values of F just derived. It should be noted that two sides of the rectangle are included in the set C and the other two sides are excluded. Thus, if there are points or line segments on the boundary of C that have positive probability, it is important to distinguish between the weak inequalities and the strict inequalities in Eq. (3.4.6).

Theorem 3.4.5 Let X and Y have a joint c.d.f. F. The c.d.f. F_1 of just the single random variable X can be derived from the joint c.d.f. F as $F_1(x) = \lim_{y\to\infty} F(x, y)$. Similarly, the c.d.f. F_2 of Y equals $F_2(y) = \lim_{x\to\infty} F(x, y)$, for $0 < y < \infty$.

Proof We prove the claim about F_1 as the claim about F_2 is similar. Let $-\infty < x < \infty$. Define

$$\begin{aligned}
B_0 &= \{X \le x \text{ and } Y \le 0\}, \\
B_n &= \{X \le x \text{ and } n-1 < Y \le n\}, \quad \text{for } n = 1, 2, \ldots, \\
A_m &= \bigcup_{n=0}^{m} B_n, \quad \text{for } m = 1, 2, \ldots.
\end{aligned}$$

Then $\{X \le x\} = \bigcup_{n=-0}^{\infty} B_n$, and $A_m = \{X \le x \text{ and } Y \le m\}$ for $m = 1, 2, \ldots$. It follows that $\Pr(A_m) = F(x, m)$ for each m. Also,

$$\begin{aligned}
F_1(x) = \Pr(X \le x) &= \Pr\left(\bigcup_{n=1}^{\infty} B_n\right) \\
&= \sum_{n=0}^{\infty} \Pr(B_n) = \lim_{m\to\infty} \Pr(A_m) \\
&= \lim_{m\to\infty} F(x, m) = \lim_{y\to\infty} F(x, y),
\end{aligned}$$

where the third equality follows from countable additivity and the fact that the B_n events are disjoint, and the last equality follows from the fact that $F(x, y)$ is monotone increasing in y for each fixed x. ■

Other relationships involving the univariate distribution of X, the univariate distribution of Y, and their joint bivariate distribution will be presented in the next section.

Finally, if X and Y have a continuous joint distribution with joint p.d.f. f, then the joint c.d.f. at (x, y) is

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(r, s)\, dr\, ds.$$

Here, the symbols r and s are used simply as dummy variables of integration. The joint p.d.f. can be derived from the joint c.d.f. by using the relations

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{\partial^2 F(x, y)}{\partial y \partial x}$$

at every point (x, y) at which these second-order derivatives exist.

Example 3.4.14

Determining a Joint p.d.f. from a Joint c.d.f. Suppose that X and Y are random variables that take values only in the intervals $0 \le X \le 2$ and $0 \le Y \le 2$. Suppose also that the joint c.d.f. of X and Y, for $0 \le x \le 2$ and $0 \le y \le 2$, is as follows:

$$F(x, y) = \frac{1}{16} xy(x + y). \tag{3.4.7}$$

We shall first determine the c.d.f. F_1 of just the random variable X and then determine the joint p.d.f. f of X and Y.

The value of $F(x, y)$ at any point (x, y) in the xy-plane that does not represent a pair of possible values of X and Y can be calculated from (3.4.7) and the fact that $F(x, y) = \Pr(X \le x \text{ and } Y \le y)$. Thus, if either $x < 0$ or $y < 0$, then $F(x, y) = 0$. If both $x > 2$ and $y > 2$, then $F(x, y) = 1$. If $0 \le x \le 2$ and $y > 2$, then $F(x, y) = F(x, 2)$, and it follows from Eq. (3.4.7) that

$$F(x, y) = \frac{1}{8} x(x + 2).$$

Similarly, if $0 \le y \le 2$ and $x > 2$, then

$$F(x, y) = \frac{1}{8} y(y + 2).$$

The function $F(x, y)$ has now been specified for every point in the xy-plane.

By letting $y \to \infty$, we find that the c.d.f. of just the random variable X is

$$F_1(x) = \begin{cases} 0 & \text{for } x < 0, \\ \frac{1}{8} x(x + 2) & \text{for } 0 \le x \le 2, \\ 1 & \text{for } x > 2. \end{cases}$$

Furthermore, for $0 < x < 2$ and $0 < y < 2$,

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{1}{8}(x + y).$$

Also, if $x < 0$, $y < 0$, $x > 2$, or $y > 2$, then

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = 0.$$

Hence, the joint p.d.f. of X and Y is

$$f(x, y) = \begin{cases} \frac{1}{8}(x + y) & \text{for } 0 < x < 2 \text{ and } 0 < y < 2, \\ 0 & \text{otherwise.} \end{cases}$$ ◀

Example 3.4.15 **Demands for Utilities.** We can compute the joint c.d.f. for water and electric demand in Example 3.4.4 by using the joint p.d.f. that was given in Eq. (3.4.2). If either $x \le 4$ or $y \le 1$, then $F(x, y) = 0$ because either $X \le x$ or $Y \le y$ would be impossible. Similarly, if both $x \ge 200$ and $y \ge 150$, $F(x, y) = 1$ because both $X \le x$ and $Y \le y$ would be sure events. For other values of x and y, we compute

$$F(x, y) = \begin{cases} \displaystyle\int_4^x \int_1^y \frac{1}{29{,}204} dy dx = \frac{xy}{29{,}204} & \text{for } 4 \le x \le 200, 1 \le y \le 150, \\ \displaystyle\int_4^x \int_1^{150} \frac{1}{29{,}204} dy dx = \frac{x}{196} & \text{for } 4 \le x \le 200, y > 150, \\ \displaystyle\int_4^{200} \int_1^y \frac{1}{29{,}204} dy dx = \frac{y}{149} & \text{for } x > 200, 1 \le y \le 150. \end{cases}$$

The reason that we need three cases in the formula for $F(x, y)$ is that the joint p.d.f. in Eq. (3.4.2) drops to 0 when x crosses above 200 or when y crosses above 150; hence, we never want to integrate 1/29,204 beyond $x = 200$ or beyond $y = 150$. If one takes the limit as $y \to \infty$ of $F(x, y)$ (for fixed $4 \le x \le 200$), one gets the second case in the formula above, which then is the c.d.f. of X, $F_1(x)$. Similarly, if one takes the $\lim_{x\to\infty} F(x, y)$ (for fixed $1 \le y \le 150$), one gets the third case in the formula, which then is the c.d.f. of Y, $F_2(y)$. ◀

Summary

The joint c.d.f. of two random variables X and Y is $F(x, y) = \Pr(X \le x \text{ and } Y \le y)$. The joint p.d.f. of two continuous random variables is a nonnegative function f such that the probability of the pair (X, Y) being in a set C is the integral of $f(x, y)$ over the set C, if the integral exists. The joint p.d.f. is also the second mixed partial derivative of the joint c.d.f. with respect to both variables. The joint p.f. of two discrete random variables is a nonnegative function f such that the probability of the pair (X, Y) being in a set C is the sum of $f(x, y)$ over all points in C. A joint p.f. can be strictly positive at countably many pairs (x, y) at most. The joint p.f./p.d.f. of a discrete random variable X and a continuous random variable Y is a nonnegative function f such that the probability of the pair (X, Y) being in a set C is obtained by summing $f(x, y)$ over all x such that $(x, y) \in C$ for each y and then integrating the resulting function of y.

Exercises

1. Suppose that the joint p.d.f. of a pair of random variables (X, Y) is constant on the rectangle where $0 \le x \le 2$ and $0 \le y \le 1$, and suppose that the p.d.f. is 0 off of this rectangle.

a. Find the constant value of the p.d.f. on the rectangle.

b. Find $\Pr(X \ge Y)$.

2. Suppose that in an electric display sign there are three light bulbs in the first row and four light bulbs in the second row. Let X denote the number of bulbs in the first row that will be burned out at a specified time t, and let Y denote the number of bulbs in the second row that will be burned out at the same time t. Suppose that the joint p.f. of X and Y is as specified in the following table:

	Y				
X	0	1	2	3	4
0	0.08	0.07	0.06	0.01	0.01
1	0.06	0.10	0.12	0.05	0.02
2	0.05	0.06	0.09	0.04	0.03
3	0.02	0.03	0.03	0.03	0.04

Determine each of the following probabilities:

a. $\Pr(X = 2)$ **b.** $\Pr(Y \ge 2)$

c. $\Pr(X \le 2 \text{ and } Y \le 2)$ **d.** $\Pr(X = Y)$

e. $\Pr(X > Y)$

3. Suppose that X and Y have a discrete joint distribution for which the joint p.f. is defined as follows:

$$f(x, y) = \begin{cases} c|x + y| & \text{for } x = -2, -1, 0, 1, 2 \text{ and} \\ & \quad y = -2, -1, 0, 1, 2, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the value of the constant c; **(b)** $\Pr(X = 0 \text{ and } Y = -2)$; **(c)** $\Pr(X = 1)$; **(d)** $\Pr(|X - Y| \le 1)$.

4. Suppose that X and Y have a continuous joint distribution for which the joint p.d.f. is defined as follows:

$$f(x, y) = \begin{cases} cy^2 & \text{for } 0 \le x \le 2 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the value of the constant c; **(b)** $\Pr(X + Y > 2)$; **(c)** $\Pr(Y < 1/2)$; **(d)** $\Pr(X \le 1)$; **(e)** $\Pr(X = 3Y)$.

5. Suppose that the joint p.d.f. of two random variables X and Y is as follows:

$$f(x, y) = \begin{cases} c(x^2 + y) & \text{for } 0 \le y \le 1 - x^2, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the value of the constant c;
(b) $\Pr(0 \le X \le 1/2)$; **(c)** $\Pr(Y \le X + 1)$;
(d) $\Pr(Y = X^2)$.

6. Suppose that a point (X, Y) is chosen at random from the region S in the xy-plane containing all points (x, y) such that $x \ge 0$, $y \ge 0$, and $4y + x \le 4$.

a. Determine the joint p.d.f. of X and Y.

b. Suppose that S_0 is a subset of the region S having area α and determine $\Pr[(X, Y) \in S_0]$.

7. Suppose that a point (X, Y) is to be chosen from the square S in the xy-plane containing all points (x, y) such that $0 \le x \le 1$ and $0 \le y \le 1$. Suppose that the probability that the chosen point will be the corner (0, 0) is 0.1, the probability that it will be the corner (1, 0) is 0.2, the probability that it will be the corner (0, 1) is 0.4, and the probability that it will be the corner (1, 1) is 0.1. Suppose also that if the chosen point is not one of the four corners of the square, then it will be an interior point of the square and will be chosen according to a constant p.d.f. over the interior of the square. Determine **(a)** $\Pr(X \le 1/4)$ and **(b)** $\Pr(X + Y \le 1)$.

8. Suppose that X and Y are random variables such that (X, Y) must belong to the rectangle in the xy-plane containing all points (x, y) for which $0 \le x \le 3$ and $0 \le y \le 4$. Suppose also that the joint c.d.f. of X and Y at every point (x, y) in this rectangle is specified as follows:

$$F(x, y) = \frac{1}{156} xy(x^2 + y).$$

Determine **(a)** $\Pr(1 \le X \le 2 \text{ and } 1 \le Y \le 2)$;
(b) $\Pr(2 \le X \le 4 \text{ and } 2 \le Y \le 4)$; **(c)** the c.d.f. of Y;
(d) the joint p.d.f. of X and Y; **(e)** $\Pr(Y \le X)$.

9. In Example 3.4.5, compute the probability that water demand X is greater than electric demand Y.

10. Let Y be the rate (calls per hour) at which calls arrive at a switchboard. Let X be the number of calls during a two-hour period. A popular choice of joint p.f./p.d.f. for (X, Y) in this example would be one like

$$f(x, y) = \begin{cases} \frac{(2y)^x}{x!} e^{-3y} & \text{if } y > 0 \text{ and } x = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

a. Verify that f is a joint p.f./p.d.f. *Hint:* First, sum over the x values using the well-known formula for the power series expansion of e^{2y}.

b. Find $\Pr(X = 0)$.

11. Consider the clinical trial of depression drugs in Example 2.1.4. Suppose that a patient is selected at random from the 150 patients in that study and we record Y, an

Table 3.3 Proportions in clinical depression study for Exercise 11

	Treatment group (Y)			
Response (X)	Imipramine (1)	Lithium (2)	Combination (3)	Placebo (4)
Relapse (0)	0.120	0.087	0.146	0.160
No relapse (1)	0.147	0.166	0.107	0.067

indicator of the treatment group for that patient, and X, an indicator of whether or not the patient relapsed. Table 3.3 contains the joint p.f. of X and Y.

a. Calculate the probability that a patient selected at random from this study used Lithium (either alone or in combination with Imipramine) and did not relapse.

b. Calculate the probability that the patient had a relapse (without regard to the treatment group).

3.5 Marginal Distributions

Earlier in this chapter, we introduced distributions for random variables, and in Sec. 3.4 we discussed a generalization to joint distributions of two random variables simultaneously. Often, we start with a joint distribution of two random variables and we then want to find the distribution of just one of them. The distribution of one random variable X computed from a joint distribution is also called the marginal distribution of X. Each random variable will have a marginal c.d.f. as well as a marginal p.d.f. or p.f. We also introduce the concept of independent random variables, which is a natural generalization of independent events.

Deriving a Marginal p.f. or a Marginal p.d.f.

We have seen in Theorem 3.4.5 that if the joint c.d.f. F of two random variables X and Y is known, then the c.d.f. F_1 of the random variable X can be derived from F. We saw an example of this derivation in Example 3.4.15. If X has a continuous distribution, we can also derive the p.d.f. of X from the joint distribution.

Example 3.5.1 Demands for Utilities. Look carefully at the formula for $F(x, y)$ in Example 3.4.15, specifically the last two branches that we identified as $F_1(x)$ and $F_2(y)$, the c.d.f.'s of the two individual random variables X and Y. It is apparent from those two formulas and Theorem 3.3.5 that the p.d.f. of X alone is

$$f_1(x) = \begin{cases} \dfrac{1}{196} & \text{for } 4 \le x \le 200, \\ 0 & \text{otherwise,} \end{cases}$$

which matches what we already found in Example 3.2.1. Similarly, the p.d.f. of Y alone is

$$f_2(y) = \begin{cases} \dfrac{1}{149} & \text{for } 1 \le y \le 150, \\ 0 & \text{otherwise.} \end{cases}$$

◀

The ideas employed in Example 3.5.1 lead to the following definition.

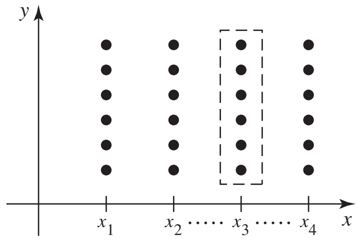


Figure 3.15 Computing $f_1(x)$ from the joint p.f.

Definition 3.5.1 Marginal c.d.f./p.f./p.d.f. Suppose that X and Y have a joint distribution. The c.d.f. of X derived by Theorem 3.4.5 is called the *marginal c.d.f.* of X. Similarly, the p.f. or p.d.f. of X associated with the marginal c.d.f. of X is called the *marginal p.f.* or *marginal p.d.f.* of X.

To obtain a specific formula for the marginal p.f. or marginal p.d.f., we start with a discrete joint distribution.

Theorem 3.5.1 If X and Y have a discrete joint distribution for which the joint p.f. is f, then the marginal p.f. f_1 of X is

$$f_1(x) = \sum_{\text{All } y} f(x, y). \tag{3.5.1}$$

Similarly, the marginal p.f. f_2 of Y is $f_2(y) = \sum_{\text{All } x} f(x, y)$.

Proof We prove the result for f_1, as the proof for f_2 is similar. We illustrate the proof in Fig. 3.15. In that figure, the set of points in the dashed box is the set of pairs with first coordinate x. The event $\{X = x\}$ can be expressed as the union of the events represented by the pairs in the dashed box, namely, $B_y = \{X = x \text{ and } Y = y\}$ for all possible y. The B_y events are disjoint and $\Pr(B_y) = f(x, y)$. Since $\Pr(X = x) = \sum_{\text{All } y} \Pr(B_y)$, Eq. (3.5.1) holds. ■

Example 3.5.2 Deriving a Marginal p.f. from a Table of Probabilities. Suppose that X and Y are the random variables in Example 3.4.3 on page 119. These are respectively the numbers of cars and televisions owned by a radomly selected household in a certain suburban area. Table 3.2 on page 119 gives their joint p.f., and we repeat that table in Table 3.4 together with row and column totals added to the margins.

The marginal p.f. f_1 of X can be read from the row totals of Table 3.4. The numbers were obtained by summing the values in each row of this table from the four columns in the central part of the table (those labeled $y = 1, 2, 3, 4$). In this way, it is found that $f_1(1) = 0.2$, $f_1(2) = 0.6$, $f_1(3) = 0.2$, and $f_1(x) = 0$ for all other values of x. This marginal p.f. gives the probabilities that a randomly selected household owns 1, 2, or 3 cars. Similarly, the marginal p.f. f_2 of Y, the probabilities that a household owns 1, 2, 3, or 4 televisions, can be read from the column totals. These numbers were obtained by adding the numbers in each of the columns from the three rows in the central part of the table (those labeled $x = 1, 2, 3$.) ◄

The name *marginal distribution* derives from the fact that the marginal distributions are the totals that appear in the margins of tables like Table 3.4.

If X and Y have a continuous joint distribution for which the joint p.d.f. is f, then the marginal p.d.f. f_1 of X is again determined in the manner shown in Eq. (3.5.1), but

Table 3.4 Joint p.f. $f(x, y)$ with marginal p.f.'s for Example 3.5.2

x	y: 1	2	3	4	Total
1	0.1	0	0.1	0	0.2
2	0.3	0	0.1	0.2	0.6
3	0	0.2	0	0	0.2
Total	0.4	0.2	0.2	0.2	1.0

the sum over all possible values of Y is now replaced by the integral over all possible values of Y.

Theorem 3.5.2 If X and Y have a continuous joint distribution with joint p.d.f. f, then the marginal p.d.f. f_1 of X is

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \quad \text{for } -\infty < x < \infty. \tag{3.5.2}$$

Similarly, the marginal p.d.f. f_2 of Y is

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx \quad \text{for } -\infty < y < \infty. \tag{3.5.3}$$

Proof We prove (3.5.2) as the proof of (3.5.3) is similar. For each x, $\Pr(X \le x)$ can be written as $\Pr((X, Y) \in C)$, where $C = \{(r, s)) : r \le x\}$. We can compute this probability directly from the joint p.d.f. of X and Y as

$$\begin{aligned} \Pr((X, Y) \in C) &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(r, s) ds dr \\ &= \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f(r, s) ds \right] dr \end{aligned} \tag{3.5.4}$$

The inner integral in the last expression of Eq. (3.5.4) is a function of r and it can easily be recognized as $f_1(r)$, where f_1 is defined in Eq. (3.5.2). It follows that $\Pr(X \le x) = \int_{-\infty}^{x} f_1(r) dr$, so f_1 is the marginal p.d.f. of X. ■

Example 3.5.3 Deriving a Marginal p.d.f. Suppose that the joint p.d.f. of X and Y is as specified in Example 3.4.8, namely,

$$f(x, y) = \begin{cases} \frac{21}{4} x^2 y & \text{for } x^2 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

The set S of points (x, y) for which $f(x, y) > 0$ is sketched in Fig. 3.16. We shall determine first the marginal p.d.f. f_1 of X and then the marginal p.d.f. f_2 of Y.

It can be seen from Fig. 3.16 that X cannot take any value outside the interval $[-1, 1]$. Therefore, $f_1(x) = 0$ for $x < -1$ or $x > 1$. Furthermore, for $-1 \le x \le 1$, it is seen from Fig. 3.16 that $f(x, y) = 0$ unless $x^2 \le y \le 1$. Therefore, for $-1 \le x \le 1$,

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{x^2}^{1} \left(\frac{21}{4}\right) x^2 y\, dy = \left(\frac{21}{8}\right) x^2 (1 - x^4).$$

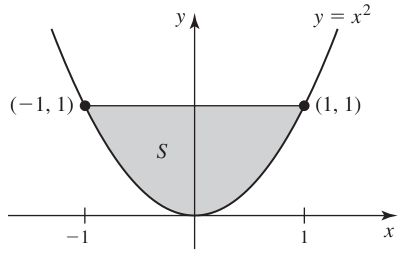


Figure 3.16 The set S where $f(x, y) > 0$ in Example 3.5.3.

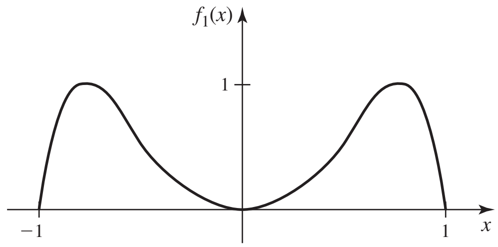


Figure 3.17 The marginal p.d.f. of X in Example 3.5.3.

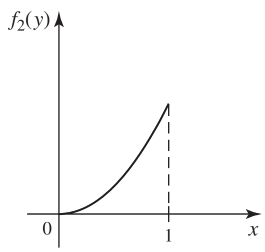


Figure 3.18 The marginal p.d.f. of Y in Example 3.5.3.

This marginal p.d.f. of X is sketched in Fig. 3.17.

Next, it can be seen from Fig. 3.16 that Y cannot take any value outside the interval [0, 1]. Therefore, $f_2(y) = 0$ for $y < 0$ or $y > 1$. Furthermore, for $0 \le y \le 1$, it is seen from Fig. 3.12 that $f(x, y) = 0$ unless $-\sqrt{y} \le x \le \sqrt{y}$. Therefore, for $0 \le y \le 1$,

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_{-\sqrt{y}}^{\sqrt{y}} \left(\frac{21}{4}\right) x^2 y\, dx = \left(\frac{7}{2}\right) y^{5/2}.$$

This marginal p.d.f. of Y is sketched in Fig. 3.18. ◀

If X has a discrete distribution and Y has a continuous distribution, we can derive the marginal p.f. of X and the marginal p.d.f. of Y from the joint p.f./p.d.f. in the same ways that we derived a marginal p.f. or a marginal p.d.f. from a joint p.f. or a joint p.d.f. The following result can be proven by combining the techniques used in the proofs of Theorems 3.5.1 and 3.5.2.

Theorem 3.5.3 Let f be the joint p.f./p.d.f. of X and Y, with X discrete and Y continuous. Then the marginal p.f. of X is

$$f_1(x) = \Pr(X = x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \quad \text{for all } x,$$

and the marginal p.d.f. of Y is

$$f_2(y) = \sum_x f(x, y), \quad \text{for } -\infty < y < \infty. \qquad \blacksquare$$

Example 3.5.4 **Determining a Marginal p.f. and Marginal p.d.f. from a Joint p.f./p.d.f.** Suppose that the joint p.f./p.d.f. of X and Y is as in Example 3.4.11 on page 124. The marginal p.f. of X is obtained by integrating

$$f_1(x) = \int_0^1 \frac{xy^{x-1}}{3} dy = \frac{1}{3},$$

for $x = 1, 2, 3$. The marginal p.d.f. of Y is obtained by summing

$$f_2(y) = \frac{1}{3} + \frac{2y}{3} + y^2, \quad \text{for } 0 < y < 1. \qquad \blacktriangleleft$$

Although the marginal distributions of X and Y can be derived from their joint distribution, it is not possible to reconstruct the joint distribution of X and Y from their marginal distributions without additional information. For instance, the marginal p.d.f.'s sketched in Figs. 3.17 and 3.18 reveal no information about the relationship between X and Y. In fact, by definition, the marginal distribution of X specifies probabilities for X without regard for the values of any other random variables. This property of a marginal p.d.f. can be further illustrated by another example.

Example 3.5.5 **Marginal and Joint Distributions.** Suppose that a penny and a nickel are each tossed n times so that every pair of sequences of tosses (n tosses in each sequence) is equally likely to occur. Consider the following two definitions of X and Y: (i) X is the number of heads obtained with the penny, and Y is the number of heads obtained with the nickel. (ii) Both X and Y are the number of heads obtained with the penny, so the random variables X and Y are actually identical.

In case (i), the marginal distribution of X and the marginal distribution of Y will be identical binomial distributions. The same pair of marginal distributions of X and Y will also be obtained in case (ii). However, the joint distribution of X and Y will not be the same in the two cases. In case (i), X and Y can take different values. Their joint p.f. is

$$f(x, y) = \begin{cases} \binom{n}{x}\binom{n}{y}\left(\frac{1}{2}\right)^{x+y} & \text{for } x = 0, 1\ldots, n,\ y = 0, 1, \ldots, n, \\ 0 & \text{otherwise.} \end{cases}$$

In case (ii), X and Y must take the same value, and their joint p.f. is

$$f(x, y) = \begin{cases} \binom{n}{x}\left(\frac{1}{2}\right)^{x} & \text{for } x = y = 0, 1\ldots, n, \\ 0 & \text{otherwise.} \end{cases} \qquad \blacktriangleleft$$

Independent Random Variables

Example 3.5.6 **Demands for Utilities.** In Examples 3.4.15 and 3.5.1, we found the marginal c.d.f.'s of water and electric demand were, respectively,

$$F_1(x) = \begin{cases} 0 & \text{for } x < 4, \\ \dfrac{x}{196} & \text{for } 4 \le x \le 200, \\ 1 & \text{for } x > 200, \end{cases} \qquad F_2(y) = \begin{cases} 0 & \text{for } y < 1, \\ \dfrac{y}{149} & \text{for } 1 \le y \le 150, \\ 1 & \text{for } y > 150. \end{cases}$$

The product of these two functions is precisely the same as the joint c.d.f. of X and Y given in Example 3.5.1. One consequence of this fact is that, for every x and y, $\Pr(X \le x,\ \text{and } Y \le y) = \Pr(X \le x)\Pr(Y \le y)$. This equation makes X and Y an example of the next definition. ◀

Definition 3.5.2 Independent Random Variables. It is said that two random variables X and Y are *independent* if, for every two sets A and B of real numbers such that $\{X \in A\}$ and $\{Y \in B\}$ are events,

$$\Pr(X \in A \text{ and } Y \in B) = \Pr(X \in A)\Pr(Y \in B). \tag{3.5.5}$$

In other words, let E be any event the occurrence or nonoccurrence of which depends only on the value of X (such as $E = \{X \in A\}$), and let D be any event the occurrence or nonoccurrence of which depends only on the value of Y (such as $D = \{Y \in B\}$). Then X and Y are independent random variables if and only if E and D are independent events for all such events E and D.

If X and Y are independent, then for all real numbers x and y, it must be true that

$$\Pr(X \le x \text{ and } Y \le y) = \Pr(X \le x)\Pr(Y \le y). \tag{3.5.6}$$

Moreover, since all probabilities for X and Y of the type appearing in Eq. (3.5.5) can be derived from probabilities of the type appearing in Eq. (3.5.6), it can be shown that if Eq. (3.5.6) is satisfied for all values of x and y, then X and Y must be independent. The proof of this statement is beyond the scope of this book and is omitted, but we summarize it as the following theorem.

Theorem 3.5.4 Let the joint c.d.f. of X and Y be F, let the marginal c.d.f. of X be F_1, and let the marginal c.d.f. of Y be F_2. Then X and Y are independent if and only if, for all real numbers x and y, $F(x, y) = F_1(x)F_2(y)$. ■

For example, the demands for water and electricity in Example 3.5.6 are independent. If one returns to Example 3.5.1, one also sees that the product of the marginal p.d.f.'s of water and electric demand equals their joint p.d.f. given in Eq. (3.4.2). This relation is characteristic of independent random variables whether discrete or continuous.

Theorem 3.5.5 Suppose that X and Y are random variables that have a joint p.f., p.d.f., or p.f./p.d.f. f. Then X and Y will be independent if and only if f can be represented in the following form for $-\infty < x < \infty$ and $-\infty < y < \infty$:

$$f(x, y) = h_1(x)h_2(y), \tag{3.5.7}$$

where h_1 is a nonnegative function of x alone and h_2 is a nonnegative function of y alone.

Proof We shall give the proof only for the case in which X is discrete and Y is continuous. The other cases are similar. For the "if" part, assume that Eq. (3.5.7) holds. Write

$$f_1(x) = \int_{-\infty}^{\infty} h_1(x)h_2(y)dy = c_1h_1(x),$$

where $c_1 = \int_{-\infty}^{\infty} h_2(y)dy$ must be finite and strictly positive, otherwise f_1 wouldn't be a p.f. So, $h_1(x) = f_1(x)/c_1$. Similarly,

$$f_2(y) = \sum_x h_1(x)h_2(y) = h_2(y)\sum_x \frac{1}{c_1}f_1(x) = \frac{1}{c_1}h_2(y).$$

So, $h_2(y) = c_1 f_2(y)$. Since $f(x, y) = h_1(x)h_2(y)$, it follows that

$$f(x, y) = \frac{f_1(x)}{c_1} c_1 f_2(y) = f_1(x) f_2(y). \tag{3.5.8}$$

Now let A and B be sets of real numbers. Assuming the integrals exist, we can write

$$\begin{aligned}\Pr(X \in A \text{ and } Y \in B) &= \sum_{x \in A} \int_B f(x, y)\,dy \\ &= \int_B \sum_{x \in A} f_1(x) f_2(y)\,dy, \\ &= \sum_{x \in A} f_1(x) \int_B f_2(y)\,dy,\end{aligned}$$

where the first equality is from Definition 3.4.5, the second is from Eq. (3.5.8), and the third is straightforward rearrangement. We now see that X and Y are independent according to Definition 3.5.2.

For the "only if" part, assume that X and Y are independent. Let A and B be sets of real numbers. Let f_1 be the marginal p.d.f. of X, and let f_2 be the marginal p.f. of Y. Then

$$\begin{aligned}\Pr(X \in A \text{ and } Y \in B) &= \sum_{x \in A} f_1(x) \int_B f_2(y)\,dy \\ &= \int_B \sum_{x \in A} f_1(x) f_2(y)\,dy,\end{aligned}$$

(if the integral exists) where the first equality follows from Definition 3.5.2 and the second is a straightforward rearrangement. We now see that $f_1(x) f_2(y)$ satisfies the conditions needed to be $f(x, y)$ as stated in Definition 3.4.5. ■

A simple corollary follows from Theorem 3.5.5.

Corollary 3.5.1 Two random variables X and Y are independent if and only if the following factorization is satisfied for all real numbers x and y:

$$f(x, y) = f_1(x) f_2(y). \tag{3.5.9}$$

■

As stated in Sec. 3.2 (see page 102), in a continuous distribution the values of a p.d.f. can be changed arbitrarily at any countable set of points. Therefore, for such a distribution it would be more precise to state that the random variables X and Y are independent if and only if it is possible to choose versions of f, f_1, and f_2 such that Eq. (3.5.9) is satisfied for $-\infty < x < \infty$ and $-\infty < y < \infty$.

The Meaning of Independence We have given a mathematical definition of independent random variables in Definition 3.5.2, but we have not yet given any interpretation of the concept of independent random variables. Because of the close connection between independent events and independent random variables, the interpretation of independent random variables should be closely related to the interpretation of independent events. We model two events as independent if learning that one of them occurs does not change the probability that the other one occurs. It is easiest to extend this idea to discrete random variables. Suppose that X and Y

Table 3.5 Joint p.f. $f(x, y)$ for Example 3.5.7

x	y = 1	2	3	4	5	6	Total
0	1/24	1/24	1/24	1/24	1/24	1/24	1/4
1	1/12	1/12	1/12	1/12	1/12	1/12	1/2
2	1/24	1/24	1/24	1/24	1/24	1/24	1/4
Total	1/6	1/6	1/6	1/6	1/6	1/6	1.000

have a discrete joint distribution. If, for each y, learning that $Y = y$ does not change any of the probabilities of the events $\{X = x\}$, we would like to say that X and Y are independent. From Corollary 3.5.1 and the definition of marginal p.f., we see that indeed X and Y are independent if and only if, for each y and x such that $\Pr(Y = y) > 0$, $\Pr(X = x|Y = y) = \Pr(X = x)$, that is, learning the value of Y doesn't change any of the probabilities associated with X. When we formally define conditional distributions in Sec. 3.6, we shall see that this interpretation of independent discrete random variables extends to all bivariate distributions. In summary, if we are trying to decide whether or not to model two random variables X and Y as independent, we should think about whether we would change the distribution of X after we learned the value of Y or vice versa.

Example 3.5.7 **Games of Chance.** A carnival game consists of rolling a fair die, tossing a fair coin two times, and recording both outcomes. Let Y stand for the number on the die, and let X stand for the number of heads in the two tosses. It seems reasonable to believe that all of the events determined by the roll of the die are independent of all of the events determined by the flips of the coin. Hence, we can assume that X and Y are independent random variables. The marginal distribution of Y is the uniform distribution on the integers $1, \dots, 6$, while the distribution of X is the binomial distribution with parameters 2 and 1/2. The marginal p.f.'s and the joint p.f. of X and Y are given in Table 3.5, where the joint p.f. was constructed using Eq. (3.5.9). The Total column gives the marginal p.f. f_1 of X, and the Total row gives the marginal p.f. f_2 of Y. ◀

Example 3.5.8 **Determining Whether Random Variables Are Independent in a Clinical Trial.** Return to the clinical trial of depression drugs in Exercise 11 of Sec. 3.4 (on page 129). In that trial, a patient is selected at random from the 150 patients in the study and we record Y, an indicator of the treatment group for that patient, and X, an indicator of whether or not the patient relapsed. Table 3.6 repeats the joint p.f. of X and Y along with the marginal distributions in the margins. We shall determine whether or not X and Y are independent.

In Eq. (3.5.9), $f(x, y)$ is the probability in the xth row and the yth column of the table, $f_1(x)$ is the number in the Total column in the xth row, and $f_2(y)$ is the number in the Total row in the yth column. It is seen in the table that $f(1, 2) = 0.087$, while $f_1(1) = 0.513$, and $f_2(1) = 0.253$. Hence, $f(1, 2) \neq f_1(1)f_2(1) = 0.129$. It follows that X and Y are not independent. ◀

It should be noted from Examples 3.5.7 and 3.5.8 that X and Y will be independent if and only if the rows of the table specifying their joint p.f. are proportional to

Table 3.6 Proportions marginals in Example 3.5.8

	Treatment group (Y)				
Response (X)	Imipramine (1)	Lithium (2)	Combination (3)	Placebo (4)	Total
Relapse (0)	0.120	0.087	0.146	0.160	0.513
No relapse (1)	0.147	0.166	0.107	0.067	0.487
Total	0.267	0.253	0.253	0.227	1.0

one another, or equivalently, if and only if the columns of the table are proportional to one another.

Example 3.5.9 Calculating a Probability Involving Independent Random Variables. Suppose that two measurements X and Y are made of the rainfall at a certain location on May 1 in two consecutive years. It might be reasonable, given knowledge of the history of rainfall on May 1, to treat the random variables X and Y as independent. Suppose that the p.d.f. g of each measurement is as follows:

$$g(x) = \begin{cases} 2x & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We shall determine the value of $\Pr(X + Y \le 1)$.

Since X and Y are independent and each has the p.d.f. g, it follows from Eq. (3.5.9) that for all values of x and y the joint p.d.f. $f(x, y)$ of X and Y will be specified by the relation $f(x, y) = g(x)g(y)$. Hence,

$$f(x, y) = \begin{cases} 4xy & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

The set S in the xy-plane, where $f(x, y) > 0$, and the subset S_0, where $x + y \le 1$, are sketched in Fig. 3.19. Thus,

$$\Pr(X + Y \le 1) = \int_{S_0}\int f(x, y)\, dx\, dy = \int_0^1 \int_0^{1-x} 4xy\, dy\, dx = \frac{1}{6}.$$

As a final note, if the two measurements X and Y had been made on the same day at nearby locations, then it might not make as much sense to treat them as independent, since we would expect them to be more similar to each other than to historical rainfalls. For example, if we first learn that X is small compared to historical rainfall on the date in question, we might then expect Y to be smaller than the historical distribution would suggest. ◄

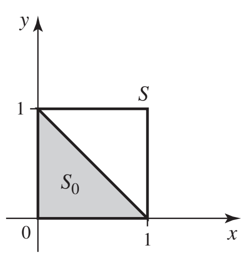


Figure 3.19 The subset S_0 where $x + y \le 1$ in Example 3.5.9.

Theorem 3.5.5 says that X and Y are independent if and only if, for all values of x and y, f can be factored into the product of an arbitrary nonnegative function of x and an arbitrary nonnegative function of y. However, it should be emphasized that, just as in Eq. (3.5.9), the factorization in Eq. (3.5.7) must be satisfied for all values of x and y ($-\infty < x < \infty$ and $-\infty < y < \infty$).

Example 3.5.10 Dependent Random Variables. Suppose that the joint p.d.f. of X and Y has the following form:

$$f(x, y) = \begin{cases} kx^2y^2 & \text{for } x^2 + y^2 \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We shall show that X and Y are not independent.

It is evident that at each point inside the circle $x^2 + y^2 \le 1$, $f(x, y)$ can be factored as in Eq. (3.5.7). However, this same factorization cannot also be satisfied at every point outside this circle. For example, $f(0.9, 0.9) = 0$, but neither $f_1(0.9) = 0$ nor $f_2(0.9) = 0$. (In Exercise 13, you can verify this feature of f_1 and f_2.)

The important feature of this example is that the values of X and Y are constrained to lie inside a circle. The joint p.d.f. of X and Y is positive inside the circle and zero outside the circle. Under these conditions, X and Y cannot be independent, because for every given value y of Y, the possible values of X will depend on y. For example, if $Y = 0$, then X can have any value such that $X^2 \le 1$; if $Y = 1/2$, then X must have a value such that $X^2 \le 3/4$. ◀

Example 3.5.10 shows that one must be careful when trying to apply Theorem 3.5.5. The situation that arose in that example will occur whenever $\{(x, y) : f(x, y) > 0\}$ has boundaries that are curved or not parallel to the coordinate axes. There is one important special case in which it is easy to check the conditions of Theorem 3.5.5. The proof is left as an exercise.

Theorem 3.5.6 Let X and Y have a continuous joint distribution. Suppose that $\{(x, y) : f(x, y) > 0\}$ is a rectangular region R (possibly unbounded) with sides (if any) parallel to the coordinate axes. Then X and Y are independent if and only if Eq. (3.5.7) holds for all $(x, y) \in R$. ■

Example 3.5.11 Verifying the Factorization of a Joint p.d.f. Suppose that the joint p.d.f. f of X and Y is as follows:

$$f(x, y) = \begin{cases} ke^{-(x+2y)} & \text{for } x \ge 0 \text{ and } y \ge 0, \\ 0 & \text{otherwise,} \end{cases}$$

where k is some constant. We shall first determine whether X and Y are independent and then determine their marginal p.d.f.'s.

In this example, $f(x, y) = 0$ outside of an unbounded rectangular region R whose sides are the lines $x = 0$ and $y = 0$. Furthermore, at each point inside R, $f(x, y)$ can be factored as in Eq. (3.5.7) by letting $h_1(x) = ke^{-x}$ and $h_2(y) = e^{-2y}$. Therefore, X and Y are independent.

It follows that in this case, except for constant factors, $h_1(x)$ for $x \ge 0$ and $h_2(y)$ for $y \ge 0$ must be the marginal p.d.f.'s of X and Y. By choosing constants that make $h_1(x)$ and $h_2(y)$ integrate to unity, we can conclude that the marginal p.d.f.'s f_1 and f_2 of X and Y must be as follows:

$$f_1(x) = \begin{cases} e^{-x} & \text{for } x \ge 0, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_2(y) = \begin{cases} 2e^{-2y} & \text{for } y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

If we multiply $f_1(x)$ times $f_2(y)$ and compare the product to $f(x, y)$, we see that $k = 2$. ◀

Note: Separate Functions of Independent Random Variables Are Independent. If X and Y are independent, then $h(X)$ and $g(Y)$ are independent no matter what the functions h and g are. This is true because for every t, the event $\{h(X) \leq t\}$ can always be written as $\{X \in A\}$, where $A = \{x : h(x) \leq t\}$. Similarly, $\{g(Y) \leq u\}$ can be written as $\{Y \in B\}$, so Eq. (3.5.6) for $h(X)$ and $g(Y)$ follows from Eq. (3.5.5) for X and Y.

Summary

Let $f(x, y)$ be a joint p.f., joint p.d.f., or joint p.f./p.d.f. of two random variables X and Y. The marginal p.f. or p.d.f. of X is denoted by $f_1(x)$, and the marginal p.f. or p.d.f. of Y is denoted by $f_2(y)$. To obtain $f_1(x)$, compute $\sum_y f(x, y)$ if Y is discrete or $\int_{-\infty}^{\infty} f(x, y)\, dy$ if Y is continuous. Similarly, to obtain $f_2(y)$, compute $\sum_x f(x, y)$ if X is discrete or $\int_{-\infty}^{\infty} f(x, y)\, dx$ if X is continuous. The random variables X and Y are independent if and only if $f(x, y) = f_1(x) f_2(y)$ for *all* x and y. This is true regardless of whether X and/or Y is continuous or discrete. A sufficient condition for two continuous random variables to be independent is that $R = \{(x, y) : f(x, y) > 0\}$ be rectangular with sides parallel to the coordinate axes and that $f(x, y)$ factors into separate functions of x of y in R.

Exercises

1. Suppose that X and Y have a continuous joint distribution for which the joint p.d.f. is

$$f(x, y) = \begin{cases} k & \text{for } a \leq x \leq b \text{ and } c \leq y \leq d, \\ 0 & \text{otherwise,} \end{cases}$$

where $a < b$, $c < d$, and $k > 0$. Find the marginal distributions of X and Y.

2. Suppose that X and Y have a discrete joint distribution for which the joint p.f. is defined as follows:

$$f(x, y) = \begin{cases} \frac{1}{30}(x + y) & \text{for } x = 0, 1, 2 \text{ and } y = 0, 1, 2, 3, \\ 0 & \text{otherwise.} \end{cases}$$

a. Determine the marginal p.f.'s of X and Y.

b. Are X and Y independent?

3. Suppose that X and Y have a continuous joint distribution for which the joint p.d.f. is defined as follows:

$$f(x, y) = \begin{cases} \frac{3}{2}y^2 & \text{for } 0 \leq x \leq 2 \text{ and } 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

a. Determine the marginal p.d.f.'s of X and Y.

b. Are X and Y independent?

c. Are the event $\{X < 1\}$ and the event $\{Y \geq 1/2\}$ independent?

4. Suppose that the joint p.d.f. of X and Y is as follows:

$$f(x, y) = \begin{cases} \frac{15}{4}x^2 & \text{for } 0 \leq y \leq 1 - x^2, \\ 0 & \text{otherwise.} \end{cases}$$

a. Determine the marginal p.d.f.'s of X and Y.

b. Are X and Y independent?

5. A certain drugstore has three public telephone booths. For $i = 0, 1, 2, 3$, let p_i denote the probability that exactly i telephone booths will be occupied on any Monday evening at 8:00 P.M.; and suppose that $p_0 = 0.1$, $p_1 = 0.2$, $p_2 = 0.4$, and $p_3 = 0.3$. Let X and Y denote the number of booths that will be occupied at 8:00 P.M. on two independent Monday evenings. Determine: **(a)** the joint p.f. of X and Y; **(b)** $\Pr(X = Y)$; **(c)** $\Pr(X > Y)$.

6. Suppose that in a certain drug the concentration of a particular chemical is a random variable with a continuous distribution for which the p.d.f. g is as follows:

$$g(x) = \begin{cases} \frac{3}{8}x^2 & \text{for } 0 \le x \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that the concentrations X and Y of the chemical in two separate batches of the drug are independent random variables for each of which the p.d.f. is g. Determine **(a)** the joint p.d.f. of X and Y; **(b)** $\Pr(X = Y)$; **(c)** $\Pr(X > Y)$; **(d)** $\Pr(X + Y \le 1)$.

7. Suppose that the joint p.d.f. of X and Y is as follows:

$$f(x, y) = \begin{cases} 2xe^{-y} & \text{for } 0 \le x \le 1 \text{ and } 0 < y < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Are X and Y independent?

8. Suppose that the joint p.d.f. of X and Y is as follows:

$$f(x, y) = \begin{cases} 24xy & \text{for } x \ge 0, y \ge 0, \text{ and } x + y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Are X and Y independent?

9. Suppose that a point (X, Y) is chosen at random from the rectangle S defined as follows:

$$S = \{(x, y) : 0 \le x \le 2 \text{ and } 1 \le y \le 4\}.$$

a. Determine the joint p.d.f. of X and Y, the marginal p.d.f. of X, and the marginal p.d.f. of Y.

b. Are X and Y independent?

10. Suppose that a point (X, Y) is chosen at random from the circle S defined as follows:

$$S = \{(x, y) : x^2 + y^2 \le 1\}.$$

a. Determine the joint p.d.f. of X and Y, the marginal p.d.f. of X, and the marginal p.d.f. of Y.

b. Are X and Y independent?

11. Suppose that two persons make an appointment to meet between 5 P.M. and 6 P.M. at a certain location, and they agree that neither person will wait more than 10 minutes for the other person. If they arrive independently at random times between 5 P.M. and 6 P.M., what is the probability that they will meet?

12. Prove Theorem 3.5.6.

13. In Example 3.5.10, verify that X and Y have the same marginal p.d.f.'s and that

$$f_1(x) = \begin{cases} 2kx^2(1 - x^2)^{3/2}/3 & \text{if } -1 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

14. For the joint p.d.f. in Example 3.4.7, determine whether or not X and Y are independent.

15. A painting process consists of two stages. In the first stage, the paint is applied, and in the second stage, a protective coat is added. Let X be the time spent on the first stage, and let Y be the time spent on the second stage. The first stage involves an inspection. If the paint fails the inspection, one must wait three minutes and apply the paint again. After a second application, there is no further inspection. The joint p.d.f. of X and Y is

$$f(x, y) = \begin{cases} \frac{1}{3} & \text{if } 1 < x < 3 \text{ and } 0 < y < 1, \\ \frac{1}{6} & \text{if } 6 < x < 8 \text{ and } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

a. Sketch the region where $f(x, y) > 0$. Note that it is not exactly a rectangle.

b. Find the marginal p.d.f.'s of X and Y.

c. Show that X and Y are independent.

This problem does not contradict Theorem 3.5.6. In that theorem the conditions, including that the set where $f(x, y) > 0$ be rectangular, are sufficient but not necessary.

3.6 Conditional Distributions

We generalize the concept of conditional probability to conditional distributions. Recall that distributions are just collections of probabilities of events determined by random variables. Conditional distributions will be the probabilities of events determined by some random variables conditional on events determined by other random variables. The idea is that there will typically be many random variables of interest in an applied problem. After we observe some of those random variables, we want to be able to adjust the probabilities associated with the ones that have not yet been observed. The conditional distribution of one random variable X given another Y will be the distribution that we would use for X after we learn the value of Y.

Table 3.7 Joint p.f. for Example 3.6.1

	Brand Y					
Stolen X	1	2	3	4	5	Total
0	0.129	0.298	0.161	0.280	0.108	0.976
1	0.010	0.010	0.001	0.002	0.001	0.024
Total	0.139	0.308	0.162	0.282	0.109	1.000

Discrete Conditional Distributions

Example 3.6.1 Auto Insurance. Insurance companies keep track of how likely various cars are to be stolen. Suppose that a company in a particular area computes the joint distribution of car brands and the indicator of whether the car will be stolen during a particular year that appears in Table 3.7.

We let $X = 1$ mean that a car is stolen, and we let $X = 0$ mean that the car is not stolen. We let Y take one of the values from 1 to 5 to indicate the brand of car as indicated in Table 3.7. If a customer applies for insurance for a particular brand of car, the company needs to compute the distribution of the random variable X as part of its premium determination. The insurance company might adjust their premium according to a risk factor such as likelihood of being stolen. Although, overall, the probability that a car will be stolen is 0.024, if we assume that we know the brand of car, the probability might change quite a bit. This section introduces the formal concepts for addressing this type of problem. ◀

Suppose that X and Y are two random variables having a discrete joint distribution for which the joint p.f. is f. As before, we shall let f_1 and f_2 denote the marginal p.f.'s of X and Y, respectively. After we observe that $Y = y$, the probability that the random variable X will take a particular value x is specified by the following conditional probability:

$$\begin{aligned}\Pr(X = x|Y = y) &= \frac{\Pr(X = x \text{ and } Y = y)}{\Pr(Y = y)} \\ &= \frac{f(x, y)}{f_2(y)}. \end{aligned} \tag{3.6.1}$$

In other words, if it is known that $Y = y$, then the probability that $X = x$ will be updated to the value in Eq. (3.6.1). Next, we consider the entire distribution of X after learning that $Y = y$.

Definition 3.6.1 Conditional Distribution/p.f. Let X and Y have a discrete joint distribution with joint p.f. f. Let f_2 denote the marginal p.f. of Y. For each y such that $f_2(y) > 0$, define

$$g_1(x|y) = \frac{f(x, y)}{f_2(y)}. \tag{3.6.2}$$

Then g_1 is called the *conditional p.f. of X given Y*. The discrete distribution whose p.f. is $g_1(\cdot|y)$ is called the *conditional distribution of X given that $Y = y$*.

Table 3.8 Conditional p.f. of Y given X for Example 3.6.3

	Brand Y				
Stolen X	1	2	3	4	5
0	0.928	0.968	0.994	0.993	0.991
1	0.072	0.032	0.006	0.007	0.009

We should verify that $g_1(x|y)$ is actually a p.f. as a function of x for each y. Let y be such that $f_2(y) > 0$. Then $g_1(x|y) \ge 0$ for all x and

$$\sum_x g_1(x|y) = \frac{1}{f_2(y)} \sum_x f(x, y) = \frac{1}{f_2(y)} f_2(y) = 1.$$

Notice that we do not bother to define $g_1(x|y)$ for those y such that $f_2(y) = 0$.

Similarly, if x is a given value of X such that $f_1(x) = \Pr(X = x) > 0$, and if $g_2(y|x)$ is the *conditional p.f. of Y given that X* $= x$, then

$$g_2(y|x) = \frac{f(x, y)}{f_1(x)}. \tag{3.6.3}$$

For each x such that $f_1(x) > 0$, the function $g_2(y|x)$ will be a p.f. as a function of y.

Example 3.6.2 Calculating a Conditional p.f. from a Joint p.f. Suppose that the joint p.f. of X and Y is as specified in Table 3.4 in Example 3.5.2. We shall determine the conditional p.f. of Y given that $X = 2$.

The marginal p.f. of X appears in the Total column of Table 3.4, so $f_1(2) = \Pr(X = 2) = 0.6$. Therefore, the conditional probability $g_2(y|2)$ that Y will take a particular value y is

$$g_2(y|2) = \frac{f(2, y)}{0.6}.$$

It should be noted that for all possible values of y, the conditional probabilities $g_2(y|2)$ must be proportional to the joint probabilities $f(2, y)$. In this example, each value of $f(2, y)$ is simply divided by the constant $f_1(2) = 0.6$ in order that the sum of the results will be equal to 1. Thus,

$$g_2(1|2) = 1/2, \quad g_2(2|2) = 0, \quad g_2(3|2) = 1/6, \quad g_2(4|2) = 1/3.$$ ◀

Example 3.6.3 Auto Insurance. Consider again the probabilities of car brands and cars being stolen in Example 3.6.1. The conditional distribution of X (being stolen) given Y (brand) is given in Table 3.8. It appears that Brand 1 is much more likely to be stolen than other cars in this area, with Brand 1 also having a significant chance of being stolen. ◀

Continuous Conditional Distributions

Example 3.6.4 Processing Times. A manufacturing process consists of two stages. The first stage takes Y minutes, and the whole process takes X minutes (which includes the first

Y minutes). Suppose that X and Y have a joint continuous distribution with joint p.d.f.

$$f(x, y) = \begin{cases} e^{-x} & \text{for } 0 \le y \le x < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

After we learn how much time Y that the first stage takes, we want to update our distribution for the total time X. In other words, we would like to be able to compute a conditional distribution for X given $Y = y$. We cannot argue the same way as we did with discrete joint distributions, because $\{Y = y\}$ is an event with probability 0 for all y. ◄

To facilitate the solutions of problems such as the one posed in Example 3.6.4, the concept of conditional probability will be extended by considering the definition of the conditional p.f. of X given in Eq. (3.6.2) and the analogy between a p.f. and a p.d.f.

Definition 3.6.2 Conditional p.d.f. Let X and Y have a continuous joint distribution with joint p.d.f. f and respective marginals f_1 and f_2. Let y be a value such that $f_2(y) > 0$. Then the *conditional p.d.f. g_1 of X given that $Y = y$* is defined as follows:

$$g_1(x|y) = \frac{f(x, y)}{f_2(y)} \quad \text{for } -\infty < x < \infty. \tag{3.6.4}$$

For values of y such that $f_2(y) = 0$, we are free to define $g_1(x|y)$ however we wish, so long as $g_1(x|y)$ is a p.d.f. as a function of x.

It should be noted that Eq. (3.6.2) and Eq. (3.6.4) are identical. However, Eq. (3.6.2) was *derived* as the conditional probability that $X = x$ given that $Y = y$, whereas Eq. (3.6.4) was *defined* to be the value of the conditional p.d.f. of X given that $Y = y$. In fact, we should verify that $g_1(x|y)$ as defined above really is a p.d.f.

Theorem 3.6.1 For each y, $g_1(x|y)$ defined in Definition 3.6.2 is a p.d.f. as a function of x.

Proof If $f_2(y) = 0$, then g_1 is defined to be any p.d.f. we wish, and hence it is a p.d.f. If $f_2(y) > 0$, g_1 is defined by Eq. (3.6.4). For each such y, it is clear that $g_1(x|y) \ge 0$ for all x. Also, if $f_2(y) > 0$, then

$$\int_{-\infty}^{\infty} g_1(x|y)\, dx = \frac{\int_{-\infty}^{\infty} f(x, y)\, dx}{f_2(y)} = \frac{f_2(y)}{f_2(y)} = 1,$$

by using the formula for $f_2(y)$ in Eq. (3.5.3). ■

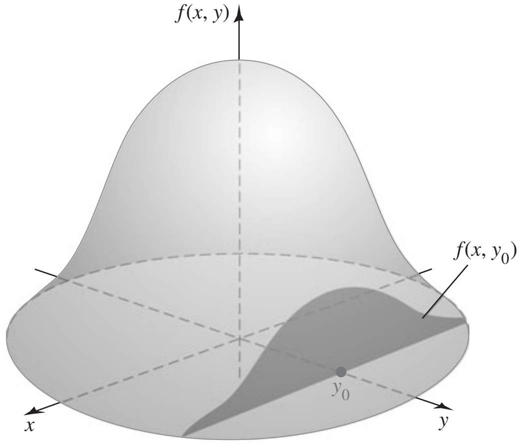
Example 3.6.5 Processing Times. In Example 3.6.4, Y is the time that the first stage of a process takes, while X is the total time of the two stages. We want to calculate the conditional p.d.f. of X given Y. We can calculate the marginal p.d.f. of Y as follows: For each y, the possible values of X are all $x \ge y$, so for each $y > 0$,

$$f_2(y) = \int_y^{\infty} e^{-x} dx = e^{-y},$$

and $f_2(y) = 0$ for $y < 0$. For each $y \ge 0$, the conditional p.d.f. of X given $Y = y$ is then

$$g_1(x|y) = \frac{f(x, y)}{f_2(y)} = \frac{e^{-x}}{e^{-y}} = e^{y-x}, \text{ for } x \ge y,$$

Figure 3.20 The conditional p.d.f. $g_1(x|y_0)$ is proportional to $f(x, y_0)$.



and $g_1(x|y) = 0$ for $x < y$. So, for example, if we observe $Y = 4$ and we want the conditional probability that $X \geq 9$, we compute

$$\Pr(X \geq 9|Y = 4) = \int_9^\infty e^{4-x} dx = e^{-5} = 0.0067. \qquad \blacktriangleleft$$

Definition 3.6.2 has an interpretation that can be understood by considering Fig. 3.20. The joint p.d.f. f defines a surface over the xy-plane for which the height $f(x, y)$ at each point (x, y) represents the relative likelihood of that point. For instance, if it is known that $Y = y_0$, then the point (x, y) must lie on the line $y = y_0$ in the xy-plane, and the relative likelihood of any point (x, y_0) on this line is $f(x, y_0)$. Hence, the conditional p.d.f. $g_1(x|y_0)$ of X should be proportional to $f(x, y_0)$. In other words, $g_1(x|y_0)$ is essentially the same as $f(x, y_0)$, but it includes a constant factor $1/[f_2(y_0)]$, which is required to make the conditional p.d.f. integrate to unity over all values of x.

Similarly, for each value of x such that $f_1(x) > 0$, the *conditional p.d.f. of Y given that $X = x$* is defined as follows:

$$g_2(y|x) = \frac{f(x, y)}{f_1(x)} \quad \text{for } -\infty < y < \infty. \tag{3.6.5}$$

This equation is identical to Eq. (3.6.3), which was derived for discrete distributions. If $f_1(x) = 0$, then $g_2(y|x)$ is arbitrary so long as it is a p.d.f. as a function of y.

Example 3.6.6 Calculating a Conditional p.d.f. from a Joint p.d.f. Suppose that the joint p.d.f. of X and Y is as specified in Example 3.4.8 on page 122. We shall first determine the conditional p.d.f. of Y given that $X = x$ and then determine some probabilities for Y given the specific value $X = 1/2$.

The set S for which $f(x, y) > 0$ was sketched in Fig. 3.12 on page 123. Furthermore, the marginal p.d.f. f_1 was derived in Example 3.5.3 on page 132 and sketched in Fig. 3.17 on page 133. It can be seen from Fig. 3.17 that $f_1(x) > 0$ for $-1 < x < 1$ but not for $x = 0$. Therefore, for each given value of x such that $-1 < x < 0$ or $0 < x < 1$, the conditional p.d.f. $g_2(y|x)$ of Y will be as follows:

$$g_2(y|x) = \begin{cases} \dfrac{2y}{1 - x^4} & \text{for } x^2 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

In particular, if it is known that $X = 1/2$, then $\Pr\left(Y \ge \frac{1}{4} \middle| X = \frac{1}{2}\right) = 1$ and

$$\Pr\left(Y \ge \frac{3}{4} \middle| X = \frac{1}{2}\right) = \int_{3/4}^{1} g_2\left(y \middle| \frac{1}{2}\right) dy = \frac{7}{15}. \blacktriangleleft$$

Note: A Conditional p.d.f. Is Not the Result of Conditioning on a Set of Probability Zero. The conditional p.d.f. $g_1(x|y)$ of X given $Y = y$ is the p.d.f. we would use for X if we were to learn that $Y = y$. This sounds as if we were conditioning on the event $\{Y = y\}$, which has zero probability if Y has a continuous distribution. Actually, for the cases we shall see in this text, the value of $g_1(x|y)$ is a limit:

$$g_1(x|y) = \lim_{\epsilon \to 0} \frac{\partial}{\partial x} \Pr(X \le x | y - \epsilon < Y \le y + \epsilon). \tag{3.6.6}$$

The conditioning event $\{y - \epsilon \le Y \le y + \epsilon\}$ in Eq. (3.6.6) has positive probability if the marginal p.d.f. of Y is positive at y. The mathematics required to make this rigorous is beyond the scope of this text. (See Exercise 11 in this section and Exercises 25 and 26 in Sec. 3.11 for results that we can prove.) Another way to think about conditioning on a continuous random variable is to notice that the conditional p.d.f.'s that we compute are typically continuous as a function of the conditioning variable. This means that conditioning on $Y = y$ or on $Y = y + \epsilon$ for small ϵ will produce nearly the same conditional distribution for X. So it does not matter much if we use $Y = y$ as a surogate for Y close to y. Nevertheless, it is important to keep in mind that the conditional p.d.f. of X given $Y = y$ is better thought of as the conditional p.d.f. of X given that Y is very close to y. This wording is awkward, so we shall not use it, but we must remember the distinction between the conditional p.d.f. and conditioning on an event with probability 0. Despite this distinction, it is still legitimate to treat Y as the constant y when dealing with the conditional distribution of X given $Y = y$.

For mixed joint distributions, we continue to use Eqs. (3.6.2) and (3.6.3) to define conditional p.f.'s and p.d.f.'s.

Definition 3.6.3 Conditional p.f. or p.d.f. from Mixed Distribution. Let X be discrete and let Y be continuous with joint p.f./p.d.f. f. Then the *conditional p.f. of X given $Y = y$* is defined by Eq. (3.6.2), and the *conditional p.d.f. of Y given $X = x$* is defined by Eq. (3.6.3).

Construction of the Joint Distribution

Example 3.6.7 Defective Parts. Suppose that a certain machine produces defective and nondefective parts, but we do not know what proportion of defectives we would find among all parts that could be produced by this machine. Let P stand for the unknown proportion of defective parts among all possible parts produced by the machine. If we were to learn that $P = p$, we might be willing to say that the parts were independent of each other and each had probability p of being defective. In other words, if we condition on $P = p$, then we have the situation described in Example 3.1.9. As in that example, suppose that we examine n parts and let X stand for the number of defectives among the n examined parts. The distribution of X, assuming that we know $P = p$, is the binomial distribution with parameters n and p. That is, we can let the binomial p.f. (3.1.4) be the conditional p.f. of X given $P = p$, namely,

$$g_1(x|p) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ for } x = 0, \ldots, n.$$

We might also believe that P has a continuous distribution with p.d.f. such as $f_2(p) = 1$ for $0 \le p \le 1$. (This means that P has the uniform distribution on the interval [0, 1].) We know that the conditional p.f. g_1 of X given $P = p$ satisfies

$$g_1(x|p) = \frac{f(x, p)}{f_2(p)},$$

where f is the joint p.f./p.d.f. of X and P. If we multiply both sides of this equation by $f_2(p)$, it follows that the joint p.f./p.d.f. of X and P is

$$f(x, p) = g_1(x|p) f_2(p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad \text{for } x = 0, \dots, n, \text{ and } 0 \le p \le 1.$$

◀

The construction in Example 3.6.7 is available in general, as we explain next.

Generalizing the Multiplication Rule for Conditional Probabilities A special case of Theorem 2.1.2, the multiplication rule for conditional probabilities, says that if A and B are two events, then $\Pr(A \cap B) = \Pr(A)\Pr(B|A)$. The following theorem, whose proof is immediate from Eqs. (3.6.4) and (3.6.5), generalizes Theorem 2.1.2 to the case of two random variables.

Theorem 3.6.2 Multiplication Rule for Distributions. Let X and Y be random variables such that X has p.f. or p.d.f. $f_1(x)$ and Y has p.f. or p.d.f. $f_2(y)$. Also, assume that the conditional p.f. or p.d.f. of X given $Y = y$ is $g_1(x|y)$ while the conditional p.f. or p.d.f. of Y given $X = x$ is $g_2(y|x)$. Then for each y such that $f_2(y) > 0$ and each x,

$$f(x, y) = g_1(x|y) f_2(y), \tag{3.6.7}$$

where f is the joint p.f., p.d.f., or p.f./p.d.f. of X and Y. Similarly, for each x such that $f_1(x) > 0$ and each y,

$$f(x, y) = f_1(x) g_2(y|x). \tag{3.6.8}$$

■

In Theorem 3.6.2, if $f_2(y_0) = 0$ for some value y_0, then it can be assumed without loss of generality that $f(x, y_0) = 0$ for all values of x. In this case, both sides of Eq. (3.6.7) will be 0, and the fact that $g_1(x|y_0)$ is not uniquely defined becomes irrelevant. Hence, Eq. (3.6.7) will be satisfied for *all* values of x and y. A similar statement applies to Eq. (3.6.8).

Example 3.6.8 Waiting in a Queue. Let X be the amount of time that a person has to wait for service in a queue. The faster the server works in the queue, the shorter should be the waiting time. Let Y stand for the rate at which the server works, which we will take to be unknown. A common choice of conditional distribution for X given $Y = y$ has conditional p.d.f. for each $y > 0$:

$$g_1(x|y) = \begin{cases} ye^{-xy} & \text{for } x \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

We shall assume that Y has a continuous distribution with p.d.f. $f_2(y) = e^{-y}$ for $y > 0$. Now we can construct the joint p.d.f. of X and Y using Theorem 3.6.2:

$$f(x, y) = g_1(x|y) f_2(y) = \begin{cases} ye^{-y(x+1)} & \text{for } x \ge 0, y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

◀

Example 3.6.9 Defective Parts. Let X be the number of defective parts in a sample of size n, and let P be the proportion of defectives among all parts, as in Example 3.6.7. The joint p.f./p.d.f of X and $P = p$ was calculated there as

$$f(x, p) = g_1(x|p)f_2(p) = \binom{n}{x} p^x(1-p)^{n-x}, \quad \text{for } x = 0, \ldots, n \text{ and } 0 \le p \le 1.$$

We could now compute the conditional p.d.f. of P given $X = x$ by first finding the marginal p.f. of X:

$$f_1(x) = \int_0^1 \binom{n}{x} p^x(1-p)^{n-x} dp, \tag{3.6.9}$$

The conditional p.d.f. of P given $X = x$ is then

$$g_2(p|x) = \frac{f(x, p)}{f_1(x)} = \frac{p^x(1-p)^{n-x}}{\int_0^1 q^x(1-q)^{n-x} dq}, \quad \text{for } 0 < p < 1. \tag{3.6.10}$$

The integral in the denominator of Eq. (3.6.10) can be tedious to calculate, but it can be found. For example, if $n = 2$ and $x = 1$, we get

$$\int_0^1 q(1-q)dq = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}.$$

In this case, $g_2(p|1) = 6p(1-p)$ for $0 \le p \le 1$. ◂

Bayes' Theorem and the Law of Total Probability for Random Variables The calculation done in Eq. (3.6.9) is an example of the generalization of the law of total probability to random variables. Also, the calculation in Eq. (3.6.10) is an example of the generalization of Bayes' theorem to random variables. The proofs of these results are straightforward and not given here.

Theorem 3.6.3 Law of Total Probability for Random Variables. If $f_2(y)$ is the marginal p.f. or p.d.f. of a random variable Y and $g_1(x|y)$ is the conditional p.f. or p.d.f. of X given $Y = y$, then the marginal p.f. or p.d.f. of X is

$$f_1(x) = \sum_y g_1(x|y)f_2(y), \tag{3.6.11}$$

if Y is discrete. If Y is continuous, the marginal p.f. or p.d.f. of X is

$$f_1(x) = \int_{-\infty}^{\infty} g_1(x|y)f_2(y)\, dy. \tag{3.6.12}$$

■

There are versions of Eqs. (3.6.11) and (3.6.12) with x and y switched and the subscripts 1 and 2 switched. These versions would be used if the joint distribution of X and Y were constructed from the conditional distribution of Y given X and the marginal distribution of X.

Theorem 3.6.4 Bayes' Theorem for Random Variables. If $f_2(y)$ is the marginal p.f. or p.d.f. of a random variable Y and $g_1(x|y)$ is the conditional p.f. or p.d.f. of X given $Y = y$, then the conditional p.f. or p.d.f. of Y given $X = x$ is

$$g_2(y|x) = \frac{g_1(x|y)f_2(y)}{f_1(x)}, \tag{3.6.13}$$

where $f_1(x)$ is obtained from Eq. (3.6.11) or (3.6.12). Similarly, the conditional p.f. or p.d.f. of X given $Y = y$ is

$$g_1(x|y) = \frac{g_2(y|x)f_1(x)}{f_2(y)}, \tag{3.6.14}$$

where $f_2(y)$ is obtained from Eq. (3.6.11) or (3.6.12) with x and y switched and with the subscripts 1 and 2 switched. ■

Example 3.6.10

Choosing Points from Uniform Distributions. Suppose that a point X is chosen from the uniform distribution on the interval [0, 1], and that after the value $X = x$ has been observed $(0 < x < 1)$, a point Y is then chosen from the uniform distribution on the interval $[x, 1]$. We shall derive the marginal p.d.f. of Y.

Since X has a uniform distribution, the marginal p.d.f. of X is as follows:

$$f_1(x) = \begin{cases} 1 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, for each value $X = x$ $(0 < x < 1)$, the conditional distribution of Y is the uniform distribution on the interval $[x, 1]$. Since the length of this interval is $1 - x$, the conditional p.d.f. of Y given that $X = x$ will be

$$g_2(y|x) = \begin{cases} \dfrac{1}{1-x} & \text{for } x < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

It follows from Eq. (3.6.8) that the joint p.d.f. of X and Y will be

$$f(x, y) = \begin{cases} \dfrac{1}{1-x} & \text{for } 0 < x < y < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.6.15}$$

Thus, for $0 < y < 1$, the value of the marginal p.d.f. $f_2(y)$ of Y will be

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_0^y \frac{1}{1-x}\, dx = -\log(1 - y). \tag{3.6.16}$$

Furthermore, since Y cannot be outside the interval $0 < y < 1$, then $f_2(y) = 0$ for $y \le 0$ or $y \ge 1$. This marginal p.d.f. f_2 is sketched in Fig. 3.21. It is interesting to note that in this example the function f_2 is unbounded.

We can also find the conditional p.d.f. of X given $Y = y$ by applying Bayes' theorem (3.6.14). The product of $g_2(y|x)$ and $f_1(x)$ was already calculated in Eq. (3.6.15).

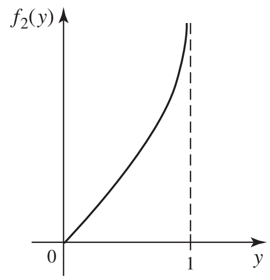


Figure 3.21 The marginal p.d.f. of Y in Example 3.6.10.

The ratio of this product to $f_2(y)$ from Eq. (3.6.16) is

$$g_1(x|y) = \begin{cases} \dfrac{-1}{(1-x)\log(1-y)} & \text{for } 0 < x < y, \\ 0 & \text{otherwise.} \end{cases} \quad \blacktriangleleft$$

Theorem 3.6.5 Independent Random Variables. Suppose that X and Y are two random variables having a joint p.f., p.d.f., or p.f./p.d.f. f. Then X and Y are independent if and only if for every value of y such that $f_2(y) > 0$ and every value of x,

$$g_1(x|y) = f_1(x). \tag{3.6.17}$$

Proof Theorem 3.5.4 says that X and Y are independent if and only if $f(x, y)$ can be factored in the following form for $-\infty < x < \infty$ and $-\infty < y < \infty$:

$$f(x, y) = f_1(x) f_2(y),$$

which holds if and only if, for all x and all y such that $f_2(y) > 0$,

$$f_1(x) = \frac{f(x, y)}{f_2(y)}. \tag{3.6.18}$$

But the right side of Eq. (3.6.18) is the formula for $g_1(x|y)$. Hence, X and Y are independent if and only if Eq. (3.6.17) holds for all x and all y such that $f_2(y) > 0$. ■

Theorem 3.6.5 says that X and Y are independent if and only if the conditional p.f. or p.d.f. of X given $Y = y$ is the same as the marginal p.f. or p.d.f. of X for all y such that $f_2(y) > 0$. Because $g_1(x|y)$ is arbitrary when $f_2(y) = 0$, we cannot expect Eq. (3.6.17) to hold in that case.

Similarly, it follows from Eq. (3.6.8) that X and Y are independent if and only if

$$g_2(y|x) = f_2(y), \tag{3.6.19}$$

for every value of x such that $f_1(x) > 0$. Theorem 3.6.5 and Eq. (3.6.19) give the mathematical justification for the meaning of independence that we presented on page 136.

Note: Conditional Distributions Behave Just Like Distributions. As we noted on page 59, conditional probabilities behave just like probabilities. Since distributions are just collections of probabilities, it follows that conditional distributions behave just like distributions. For example, to compute the conditional probability that a discrete random variable X is in some interval $[a, b]$ given $Y = y$, we must add $g_1(x|y)$ for all values of x in the interval. Also, theorems that we have proven or shall prove about distributions will have versions conditional on additional random variables. We shall postpone examples of such theorems until Sec. 3.7 because they rely on joint distributions of more than two random variables.

Summary

The conditional distribution of one random variable X given an observed value y of another random variable Y is the distribution we would use for X if we were to learn that $Y = y$. When dealing with the conditional distribution of X given $Y = y$, it is safe to behave as if Y were the constant y. If X and Y have joint p.f., p.d.f., or p.f./p.d.f. $f(x, y)$, then the conditional p.f. or p.d.f. of X given $Y = y$ is $g_1(x|y) =$

$f(x, y)/f_2(y)$, where f_2 is the marginal p.f. or p.d.f. of Y. When it is convenient to specify a conditional distribution directly, the joint distribution can be constructed from the conditional together with the other marginal. For example,

$$f(x, y) = g_1(x|y)f_2(y) = f_1(x)g_2(y|x).$$

In this case, we have versions of the law of total probability and Bayes' theorem for random variables that allow us to calculate the other marginal and conditional.

Two random variables X and Y are independent if and only if the conditional p.f. or p.d.f. of X given $Y = y$ is the same as the marginal p.f. or p.d.f. of X for all y such that $f_2(y) > 0$. Equivalently, X and Y are independent if and only if the conditional p.f. of p.d.f. of Y given $X = x$ is the same as the marginal p.f. or p.d.f. of Y for all x such that $f_1(x) > 0$.

Exercises

1. Suppose that two random variables X and Y have the joint p.d.f. in Example 3.5.10 on page 139. Compute the conditional p.d.f. of X given $Y = y$ for each y.

2. Each student in a certain high school was classified according to her year in school (freshman, sophomore, junior, or senior) and according to the number of times that she had visited a certain museum (never, once, or more than once). The proportions of students in the various classifications are given in the following table:

	Never	Once	More than once
Freshmen	0.08	0.10	0.04
Sophomores	0.04	0.10	0.04
Juniors	0.04	0.20	0.09
Seniors	0.02	0.15	0.10

a. If a student selected at random from the high school is a junior, what is the probability that she has never visited the museum?

b. If a student selected at random from the high school has visited the museum three times, what is the probability that she is a senior?

3. Suppose that a point (X, Y) is chosen at random from the disk S defined as follows:

$$S = \{(x, y) : (x - 1)^2 + (y + 2)^2 \le 9\}.$$

Determine **(a)** the conditional p.d.f. of Y for every given value of X, and **(b)** $\Pr(Y > 0|X = 2)$.

4. Suppose that the joint p.d.f. of two random variables X and Y is as follows:

$$f(x, y) = \begin{cases} c(x + y^2) & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the conditional p.d.f. of X for every given value of Y, and **(b)** $\Pr(X < \frac{1}{2}|Y = \frac{1}{2})$.

5. Suppose that the joint p.d.f. of two points X and Y chosen by the process described in Example 3.6.10 is as given by Eq. (3.6.15). Determine **(a)** the conditional p.d.f. of X for every given value of Y, and **(b)** $\Pr\left(X > \frac{1}{2}\middle|Y = \frac{3}{4}\right)$.

6. Suppose that the joint p.d.f. of two random variables X and Y is as follows:

$$f(x, y) = \begin{cases} c \sin x & \text{for } 0 \le x \le \pi/2 \text{ and } 0 \le y \le 3, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the conditional p.d.f. of Y for every given value of X, and **(b)** $\Pr(1 < Y < 2|X = 0.73)$.

7. Suppose that the joint p.d.f. of two random variables X and Y is as follows:

$$f(x, y) = \begin{cases} \frac{3}{16}(4 - 2x - y) & \text{for } x > 0,\ y > 0, \\ & \quad \text{and } 2x + y < 4, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the conditional p.d.f. of Y for every given value of X, and **(b)** $\Pr(Y \ge 2|X = 0.5)$.

8. Suppose that a person's score X on a mathematics aptitude test is a number between 0 and 1, and that his score Y on a music aptitude test is also a number between 0 and 1. Suppose further that in the population of all college students in the United States, the scores X and Y are distributed according to the following joint p.d.f.:

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y) & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

a. What proportion of college students obtain a score greater than 0.8 on the mathematics test?

b. If a student's score on the music test is 0.3, what is the probability that his score on the mathematics test will be greater than 0.8?

c. If a student's score on the mathematics test is 0.3, what is the probability that his score on the music test will be greater than 0.8?

9. Suppose that either of two instruments might be used for making a certain measurement. Instrument 1 yields a measurement whose p.d.f. h_1 is

$$h_1(x) = \begin{cases} 2x & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Instrument 2 yields a measurement whose p.d.f. h_2 is

$$h_2(x) = \begin{cases} 3x^2 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that one of the two instruments is chosen at random and a measurement X is made with it.

a. Determine the marginal p.d.f. of X.

b. If the value of the measurement is $X = 1/4$, what is the probability that instrument 1 was used?

10. In a large collection of coins, the probability X that a head will be obtained when a coin is tossed varies from one coin to another, and the distribution of X in the collection is specified by the following p.d.f.:

$$f_1(x) = \begin{cases} 6x(1-x) & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that a coin is selected at random from the collection and tossed once, and that a head is obtained. Determine the conditional p.d.f. of X for this coin.

11. The definition of the conditional p.d.f. of X given $Y = y$ is arbitrary if $f_2(y) = 0$. The reason that this causes no serious problem is that it is highly unlikely that we will observe Y close to a value y_0 such that $f_2(y_0) = 0$. To be more precise, let $f_2(y_0) = 0$, and let $A_0 = [y_0 - \epsilon, y_0 + \epsilon]$. Also, let y_1 be such that $f_2(y_1) > 0$, and let $A_1 = [y_1 - \epsilon, y_1 + \epsilon]$. Assume that f_2 is continuous at both y_0 and y_1. Show that

$$\lim_{\epsilon \to 0} \frac{\Pr(Y \in A_0)}{\Pr(Y \in A_1)} = 0.$$

That is, the probability that Y is close to y_0 is much smaller than the probability that Y is close to y_1.

12. Let Y be the rate (calls per hour) at which calls arrive at a switchboard. Let X be the number of calls during a two-hour period. Suppose that the marginal p.d.f. of Y is

$$f_2(y) = \begin{cases} e^{-y} & \text{if } y > 0, \\ 0 & \text{otherwise,} \end{cases}$$

and that the conditional p.f. of X given $Y = y$ is

$$g_1(x|y) = \begin{cases} \dfrac{(2y)^x}{x!} e^{-2y} & \text{if } x = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

a. Find the marginal p.f. of X. (You may use the formula $\int_0^\infty y^k e^{-y}\, dy = k!$.)

b. Find the conditional p.d.f. $g_2(y|0)$ of Y given $X = 0$.

c. Find the conditional p.d.f. $g_2(y|1)$ of Y given $X = 1$.

d. For what values of y is $g_2(y|1) > g_2(y|0)$? Does this agree with the intuition that the more calls you see, the higher you should think the rate is?

13. Start with the joint distribution of treatment group and response in Table 3.6 on page 138. For each treatment group, compute the conditional distribution of response given the treatment group. Do they appear to be very similar or quite different?

3.7 Multivariate Distributions

In this section, we shall extend the results that were developed in Sections 3.4, 3.5, and 3.6 for two random variables X and Y to an arbitrary finite number n of random variables $X_1, \ldots, X_n$. In general, the joint distribution of more than two random variables is called a multivariate distribution. *The theory of statistical inference (the subject of the part of this book beginning with Chapter 7) relies on mathematical models for observable data in which each observation is a random variable. For this reason, multivariate distributions arise naturally in the mathematical models for data. The most commonly used model will be one in which the individual data random variables are conditionally independent given one or two other random variables.*

Joint Distributions

Example 3.7.1 A Clinical Trial. Suppose that m patients with a certain medical condition are given a treatment, and each patient either recovers from the condition or fails to recover. For each $i = 1, \ldots, m$, we can let $X_i = 1$ if patient i recovers and $X_i = 0$ if not. We might also believe that there is a random variable P having a continuous distribution taking values between 0 and 1 such that, if we knew that $P = p$, we would say that the m patients recover or fail to recover independently of each other each with probability p of recovery. We now have named $n = m + 1$ random variables in which we are interested. ◀

The situation described in Example 3.7.1 requires us to construct a joint distribution for n random variables. We shall now provide definitions and examples of the important concepts needed to discuss multivariate distributions.

Definition 3.7.1 Joint Distribution Function/c.d.f. The *joint c.d.f.* of n random variables $X_1, \ldots, X_n$ is the function F whose value at every point $(x_1, \ldots, x_n)$ in n-dimensional space R^n is specified by the relation

$$F(x_1, \ldots, x_n) = \Pr(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n). \tag{3.7.1}$$

Every multivariate c.d.f. satisfies properties similar to those given earlier for univariate and bivariate c.d.f.'s.

Example 3.7.2 Failure Times. Suppose that a machine has three parts, and part i will fail at time X_i for $i = 1, 2, 3$. The following function might be the joint c.d.f. of X_1, X_2, and X_3:

$$F(x_1, x_2, x_3) = \begin{cases} (1 - e^{-x_1})(1 - e^{-2x_2})(1 - e^{-3x_3}) & \text{for } x_1, x_2, x_3 \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$ ◀

Vector Notation In the study of the joint distribution of n random variables $X_1, \ldots, X_n$, it is often convenient to use the vector notation $\boldsymbol{X} = (X_1, \ldots, X_n)$ and to refer to $\boldsymbol{X}$ as a *random vector*. Instead of speaking of the joint distribution of the random variables $X_1, \ldots, X_n$ with a joint c.d.f. $F(x_1, \ldots, x_n)$, we can simply speak of the distribution of the random vector $\boldsymbol{X}$ with c.d.f. $F(\boldsymbol{x})$. When this vector notation is used, it must be kept in mind that if $\boldsymbol{X}$ is an n-dimensional random vector, then its c.d.f. is defined as a function on n-dimensional space R^n. At each point $\boldsymbol{x} = (x_1, \ldots, x_n) \in R^n$, the value of $F(\boldsymbol{x})$ is specified by Eq. (3.7.1).

Definition 3.7.2 Joint Discrete Distribution/p.f. It is said that n random variables $X_1, \ldots, X_n$ have a *discrete joint distribution* if the random vector $(X_1, \ldots, X_n)$ can have only a finite number or an infinite sequence of different possible values $(x_1, \ldots, x_n)$ in R^n. The *joint p.f.* of $X_1, \ldots, X_n$ is then defined as the function f such that for every point $(x_1, \ldots, x_n) \in R^n$,

$$f(x_1, \ldots, x_n) = \Pr(X_1 = x_1, \ldots, X_n = x_n).$$

In vector notation, Definition 3.7.2 says that the random vector $\boldsymbol{X}$ has a discrete distribution and that its p.f. is specified at every point $\boldsymbol{x} \in R^n$ by the relation

$$f(\boldsymbol{x}) = \Pr(\boldsymbol{X} = \boldsymbol{x}).$$

The following result is a simple generalization of Theorem 3.4.2.

Theorem 3.7.1 If $\boldsymbol{X}$ has a joint discrete distribution with joint p.f. f, then for every subset $C \subset R^n$,

$$\Pr(\boldsymbol{X} \in C) = \sum_{x \in C} f(\boldsymbol{x}). \quad \blacksquare$$

It is easy to show that, if each of $X_1, \ldots, X_n$ has a discrete distribution, then $\boldsymbol{X} = (X_1, \ldots, X_n)$ has a discrete joint distribution.

Example 3.7.3 A Clinical Trial. Consider the m patients in Example 3.7.1. Suppose for now that $P = p$ is known so that we don't treat it as a random variable. The joint p.f. of $\boldsymbol{X} = (X_1, \ldots, X_m)$ is

$$f(\boldsymbol{x}) = p^{x_1+\cdots+x_m}(1-p)^{m-x_1-\cdots-x_m},$$

for all $x_i \in \{0, 1\}$ and 0 otherwise. ◀

Definition 3.7.3 Continuous Distribution/p.d.f. It is said that n random variables $X_1, \ldots, X_n$ have a continuous joint distribution if there is a nonnegative function f defined on R^n such that for every subset $C \subset R^n$,

$$\Pr[(X_1, \ldots, X_n) \in C] = \int \underset{C}{\cdots} \int f(x_1, \ldots x_n)\, dx_1 \cdots dx_n, \tag{3.7.2}$$

if the integral exists. The function f is called the *joint p.d.f.* of $X_1, \ldots, X_n$.

In vector notation, $f(\boldsymbol{x})$ denotes the p.d.f. of the random vector $\boldsymbol{X}$ and Eq. (3.7.2) could be rewritten more simply in the form

$$\Pr(\boldsymbol{X} \in C) = \int \underset{C}{\cdots} \int f(\boldsymbol{x})\, d\boldsymbol{x}.$$

Theorem 3.7.2 If the joint distribution of $X_1, \ldots, X_n$ is continuous, then the joint p.d.f. f can be derived from the joint c.d.f. F by using the relation

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}$$

at all points $(x_1, \ldots, x_n)$ at which the derivative in this relation exists. ■

Example 3.7.4 Failure Times. We can find the joint p.d.f. for the three random variables in Example 3.7.2 by applying Theorem 3.7.2. The third-order mixed partial is easily calculated to be

$$f(x_1, x_2, x_3) = \begin{cases} 6e^{-x_1-2x_2-3x_3} & \text{for } x_1, x_2, x_3 > 0, \\ 0 & \text{otherwise.} \end{cases}$$ ◀

It is important to note that, even if each of $X_1, \ldots, X_n$ has a continuous distribution, the vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ might not have a continuous joint distribution. See Exercise 9 in this section.

Example 3.7.5 Service Times in a Queue. A queue is a system in which customers line up for service and receive their service according to some algorithm. A simple model is the single-server queue, in which all customers wait for a single server to serve everyone ahead of them in the line and then they get served. Suppose that n customers arrive at a

single-server queue for service. Let X_i be the time that the server spends serving customer i for $i = 1, \ldots, n$. We might use a joint distribution for $\boldsymbol{X} = (X_1, \ldots, X_n)$ with joint p.d.f. of the form

$$f(\boldsymbol{x}) = \begin{cases} \dfrac{c}{\left(2 + \sum_{i=1}^{n} x_i\right)^{n+1}} & \text{for all } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.3}$$

We shall now find the value of c such that the function in Eq. (3.7.3) is a joint p.d.f. We can do this by integrating over each variable $x_1, \ldots, x_n$ in succession (starting with x_n). The first integral is

$$\int_0^\infty \frac{c}{(2 + x_1 + \cdots + x_n)^{n+1}}\, dx_n = \frac{c/n}{(2 + x_1 + \cdots + x_{n-1})^n}. \tag{3.7.4}$$

The right-hand side of Eq. (3.7.4) is in the same form as the original p.d.f. except that n has been reduced to $n - 1$ and c has been divided by n. It follows that when we integrate over the variable x_i (for $i = n - 1, n - 2, \ldots, 1$), the result will be in the same form with n reduced to $i - 1$ and c divided by $n(n - 1) \cdots i$. The result of integrating all coordinates except x_1 is then

$$\frac{c/n!}{(2 + x_1)^2}, \quad \text{for } x_1 > 0.$$

Integrating x_1 out of this yields $c/[2(n!)]$, which must equal 1, so $c = 2(n!)$. ◀

Mixed Distributions

Example 3.7.6

Arrivals at a Queue. In Example 3.7.5, we introduced the single-server queue and discussed service times. Some features that influence the performance of a queue are the rate at which customers arrive and the rate at which customers are served. Let Z stand for the rate at which customers are served, and let Y stand for the rate at which customers arrive at the queue. Finally, let W stand for the number of customers that arrive during one day. Then W is discrete while Y and Z could be continuous random variables. A possible joint p.f./p.d.f. for these three random variables is

$$f(y, z, w) = \begin{cases} 6e^{-3z-10y}(8y)^w/w! & \text{for } z, y > 0 \text{ and } w = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

We can verify this claim shortly. ◀

Definition 3.7.4

Joint p.f./p.d.f. Let $X_1, \ldots, X_n$ be random variables, some of which have a continuous joint distribution and some of which have discrete distributions; their joint distribution would then be represented by a function f that we call the *joint p.f./p.d.f*. The function has the property that the probability that $\boldsymbol{X}$ lies in a subset $C \subset R^n$ is calculated by summing $f(\boldsymbol{x})$ over the values of the coordinates of $\boldsymbol{x}$ that correspond to the discrete random variables and integrating over those coordinates that correspond to the continuous random variables for all points $\boldsymbol{x} \in C$.

Example 3.7.7

Arrivals at a Queue. We shall now verify that the proposed p.f./p.d.f. in Example 3.7.6 actually sums and integrates to 1 over all values of (y, z, w). We must sum over w and integrate over y and z. We have our choice of in what order to do them. It is not

difficult to see that we can factor f as $f(y, z, w) = h_2(z)h_{13}(y, w)$, where

$$h_2(z) = \begin{cases} 6e^{-3z} & \text{for } z > 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$h_{13}(y, w) = \begin{cases} e^{-10y}(8y)^w/w! & \text{for } y > 0 \text{ and } w = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

So we can integrate z out first to get

$$\int_{-\infty}^{\infty} f(y, z, w)dz = h_{13}(y, w) \int_0^{\infty} 6e^{-3z}dz = 2h_{13}(y, w).$$

Integrating y out of $h_{13}(y, w)$ is possible, but not pleasant. Instead, notice that $(8y)^w/w!$ is the wth term in the Taylor expansion of e^{8y}. Hence,

$$\sum_{w=0}^{\infty} 2h_{13}(y, w) = 2e^{-10y} \sum_{w=0}^{\infty} \frac{(8y)^w}{w!} = 2e^{-10y}e^{8y} = 2e^{-2y},$$

for $y > 0$ and 0 otherwise. Finally, integrating over y yields 1. ◀

Example 3.7.8 A Clinical Trial. In Example 3.7.1, one of the random variables P has a continuous distribution, and the others $X_1, \ldots, X_m$ have discrete distributions. A possible joint p.f./p.d.f. for $(X_1, \ldots, X_m, P)$ is

$$f(\boldsymbol{x}, p) = \begin{cases} p^{x_1+\cdots+x_m}(1-p)^{m-x_1-\cdots-x_m} & \text{for all } x_i \in \{0, 1\} \text{ and } 0 \le p \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We can find probabilities based on this function. Suppose, for example, that we want the probability that there is exactly one success among the first two patients, that is, $\Pr(X_1 + X_2 = 1)$. We must integrate $f(\boldsymbol{x}, p)$ over p and sum over all values of $\boldsymbol{x}$ that have $x_1 + x_2 = 1$. For purposes of illustration, suppose that $m = 4$. First, factor out $p^{x_1+x_2}(1-p)^{2-x_1-x_2} = p(1-p)$, which yields

$$f(\boldsymbol{x}, p) = [p(1-p)]p^{x_3+x_4}(1-p)^{2-x_3-x_4},$$

for $x_3, x_4 \in \{0, 1\}$, $0 < p < 1$, and $x_1 + x_2 = 1$. Summing over x_3 yields

$$[p(1-p)]\left(p^{x_4}(1-p)^{1-x_4}(1-p) + pp^{x_4}(1-p)^{1-x_4}\right) = [p(1-p)]p^{x_4}(1-p)^{1-x_4}.$$

Summing this over x_4 gives $p(1-p)$. Next, integrate over p to get $\int_0^1 p(1-p)dp = 1/6$. Finally, note that there are two (x_1, x_2) vectors, $(1, 0)$ and $(0, 1)$, that have $x_1 + x_2 = 1$, so $\Pr(X_1 + X_2 = 1) = (1/6) + (1/6) = 1/3$. ◀

Marginal Distributions

Deriving a Marginal p.d.f. If the joint distribution of n random variables $X_1, \ldots, X_n$ is known, then the marginal distribution of each single random variable X_i can be derived from this joint distribution. For example, if the joint p.d.f. of $X_1, \ldots, X_n$ is f, then the marginal p.d.f. f_1 of X_1 is specified at every value x_1 by the relation

$$f_1(x_1) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-1} f(x_1, \ldots, x_n)\, dx_2 \cdots dx_n.$$

More generally, the marginal joint p.d.f. of any k of the n random variables $X_1, \ldots, X_n$ can be found by integrating the joint p.d.f. over all possible values of

the other $n - k$ variables. For example, if f is the joint p.d.f. of four random variables X_1, X_2, X_3, and X_4, then the marginal bivariate p.d.f. f_{24} of X_2 and X_4 is specified at each point (x_2, x_4) by the relation

$$f_{24}(x_2, x_4) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4)\, dx_1\, dx_3.$$

Example 3.7.9

Service Times in a Queue. Suppose that $n = 5$ in Example 3.7.5 and that we want the marginal bivariate p.d.f. of (X_1, X_4). We must integrate Eq. (3.7.3) over x_2, x_3, and x_5. Since the joint p.d.f. is symmetric with respect to permutations of the coordinates of $\boldsymbol{x}$, we shall just integrate over the last three variables and then change the names of the remaining variables to x_1 and x_4. We already saw how to do this in Example 3.7.5. The result is

$$f_{12}(x_1, x_2) = \begin{cases} \dfrac{4}{(2 + x_1 + x_2)^3} & \text{for } x_1, x_2 > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.5}$$

Then f_{14} is just like (3.7.5) with all the 2 subscripts changed to 4. The univariate marginal p.d.f. of each X_i is

$$f_i(x_i) = \begin{cases} \dfrac{2}{(2 + x_i)^2} & \text{for } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.6}$$

So, for example, if we want to know how likely it is that a customer will have to wait longer than three time units, we can calculate $\Pr(X_i > 3)$ by integrating the function in Eq. (3.7.6) from 3 to ∞. The result is 0.4. ◀

If n random variables $X_1, \ldots, X_n$ have a discrete joint distribution, then the marginal joint p.f. of each subset of the n variables can be obtained from relations similar to those for continuous distributions. In the new relations, the integrals are replaced by sums.

Deriving a Marginal c.d.f. Consider now a joint distribution for which the joint c.d.f. of $X_1, \ldots, X_n$ is F. The marginal c.d.f. F_1 of X_1 can be obtained from the following relation:

$$\begin{aligned} F_1(x_1) &= \Pr(X_1 \le x_1) = \Pr(X_1 \le x_1, X_2 < \infty, \ldots, X_n < \infty) \\ &= \lim_{x_2, \ldots, x_n \to \infty} F(x_1, x_2, \ldots, x_n). \end{aligned}$$

Example 3.7.10

Failure Times. We can find the marginal c.d.f. of X_1 from the joint c.d.f. in Example 3.7.2 by letting x_2 and x_3 go to ∞. The limit is $F_1(x_1) = 1 - e^{-x_1}$ for $x_1 \ge 0$ and 0 otherwise. ◀

More generally, the marginal joint c.d.f. of any k of the n random variables $X_1, \ldots, X_n$ can be found by computing the limiting value of the n-dimensional c.d.f. F as $x_j \to \infty$ for each of the other $n - k$ variables x_j. For example, if F is the joint c.d.f. of four random variables X_1, X_2, X_3, and X_4, then the marginal bivariate c.d.f. F_{24} of X_2 and X_4 is specified at every point (x_2, x_4) by the relation

$$F_{24}(x_2, x_4) = \lim_{x_1, x_3 \to \infty} F(x_1, x_2, x_3, x_4).$$

Example 3.7.11 Failure Times. We can find the marginal bivariate c.d.f. of X_1 and X_3 from the joint c.d.f. in Example 3.7.2 by letting x_2 go to ∞. The limit is

$$F_{13}(x_1, x_3) = \begin{cases} (1 - e^{-x_1})(1 - e^{-3x_3}) & \text{for } x_1, x_3 \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$ ◄

Independent Random Variables

Definition 3.7.5 Independent Random Variables. It is said that n random variables $X_1, \ldots, X_n$ are *independent* if, for every n sets $A_1, A_2, \ldots, A_n$ of real numbers,

$$\Pr(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) \\ = \Pr(X_1 \in A_1)\Pr(X_2 \in A_2) \cdots \Pr(X_n \in A_n).$$

If $X_1, \ldots, X_n$ are independent, it follows easily that the random variables in every nonempty subset of $X_1, \ldots, X_n$ are also independent. (See Exercise 11.)

There is a generalization of Theorem 3.5.4.

Theorem 3.7.3 Let F denote the joint c.d.f. of $X_1, \ldots, X_n$, and let F_i denote the marginal univariate c.d.f. of X_i for $i = 1, \ldots, n$. The variables $X_1, \ldots, X_n$ are independent if and only if, for all points $(x_1, x_2, \ldots, x_n) \in R^n$,

$$F(x_1, x_2, \ldots, x_n) = F_1(x_1)F_2(x_2) \cdots F_n(x_n).$$ ■

Theorem 3.7.3 says that $X_1, \ldots, X_n$ are independent if and only if their joint c.d.f. is the product of their n individual marginal c.d.f.'s. It is easy to check that the three random variables in Example 3.7.2 are independent using Theorem 3.7.3.

There is also a generalization of Corollary 3.5.1.

Theorem 3.7.4 If $X_1, \ldots, X_n$ have a continuous, discrete, or mixed joint distribution for which the joint p.d.f., joint p.f., or joint p.f./p.d.f. is f, and if f_i is the marginal univariate p.d.f. or p.f. of X_i $(i = 1, \ldots, n)$, then $X_1, \ldots, X_n$ are independent if and only if the following relation is satisfied at all points $(x_1, x_2, \ldots, x_n) \in R^n$:

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2) \cdots f_n(x_n). \tag{3.7.7}$$ ■

Example 3.7.12 Service Times in a Queue. In Example 3.7.9, we can multiply together the two univariate marginal p.d.f.'s of X_1 and X_2 calculated using Eq. (3.7.6) and see that the product does *not* equal the bivariate marginal p.d.f. of (X_1, X_2) in Eq. (3.7.5). So X_1 and X_2 are not independent. ◄

Definition 3.7.6 Random Samples/i.i.d./Sample Size. Consider a given probability distribution on the real line that can be represented by either a p.f. or a p.d.f. f. It is said that n random variables $X_1, \ldots, X_n$ form a *random sample* from this distribution if these random variables are independent and the marginal p.f. or p.d.f. of each of them is f. Such random variables are also said to be *independent and identically distributed*, abbreviated *i.i.d.* We refer to the number n of random variables as the *sample size*.

Definition 3.7.6 says that $X_1, \ldots, X_n$ form a random sample from the distribution represented by f if their joint p.f. or p.d.f. g is specified as follows at all points $(x_1, x_2, \ldots, x_n) \in R^n$:

$$g(x_1, \ldots, x_n) = f(x_1)f(x_2) \cdots f(x_n).$$

Clearly, an i.i.d. sample cannot have a mixed joint distribution.

Example 3.7.13 Lifetimes of Light Bulbs. Suppose that the lifetime of each light bulb produced in a certain factory is distributed according to the following p.d.f.:

$$f(x) = \begin{cases} xe^{-x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

We shall determine the joint p.d.f. of the lifetimes of a random sample of n light bulbs drawn from the factory's production.

The lifetimes $X_1, \ldots, X_n$ of the selected bulbs will form a random sample from the p.d.f. f. For typographical simplicity, we shall use the notation $\exp(v)$ to denote the exponential e^v when the expression for v is complicated. Then the joint p.d.f. g of $X_1, \ldots, X_n$ will be as follows: If $x_i > 0$ for $i = 1, \ldots, n$,

$$\begin{aligned} g(x_1, \ldots, x_n) &= \prod_{i=1}^{n} f(x_i) \\ &= \left(\prod_{i=1}^{n} x_i\right) \exp\left(-\sum_{i=1}^{n} x_i\right). \end{aligned}$$

Otherwise, $g(x_1, \ldots, x_n) = 0$.

Every probability involving the n lifetimes $X_1, \ldots, X_n$ can in principle be determined by integrating this joint p.d.f. over the appropriate subset of R^n. For example, if C is the subset of points $(x_1, \ldots, x_n)$ such that $x_i > 0$ for $i = 1, \ldots, n$ and $\sum_{i=1}^{n} x_i < a$, where a is a given positive number, then

$$\Pr\left(\sum_{i=1}^{n} X_i < a\right) = \int \underset{C}{\cdots} \int \left(\prod_{i=1}^{n} x_i\right) \exp\left(-\sum_{i=1}^{n} x_i\right) dx_1 \cdots dx_n. \quad ◀$$

The evaluation of the integral given at the end of Example 3.7.13 may require a considerable amount of time without the aid of tables or a computer. Certain other probabilities, however, can be evaluated easily from the basic properties of continuous distributions and random samples. For example, suppose that for the conditions of Example 3.7.13 it is desired to find $\Pr(X_1 < X_2 < \cdots < X_n)$. Since the random variables $X_1, \ldots, X_n$ have a continuous joint distribution, the probability that at least two of these random variables will have the same value is 0. In fact, the probability is 0 that the vector $(X_1, \ldots, X_n)$ will belong to each specific subset of R^n for which the n-dimensional volume is 0. Furthermore, since $X_1, \ldots, X_n$ are independent and identically distributed, each of these variables is equally likely to be the smallest of the n lifetimes, and each is equally likely to be the largest. More generally, if the lifetimes $X_1, \ldots, X_n$ are arranged in order from the smallest to the largest, each particular ordering of $X_1, \ldots, X_n$ is as likely to be obtained as any other ordering. Since there are $n!$ different possible orderings, the probability that the particular ordering $X_1 < X_2 < \cdots < X_n$ will be obtained is $1/n!$. Hence,

$$\Pr(X_1 < X_2 < \cdots < X_n) = \frac{1}{n!}.$$

Conditional Distributions

Suppose that n random variables $X_1, \ldots, X_n$ have a continuous joint distribution for which the joint p.d.f. is f and that f_0 denotes the marginal joint p.d.f. of the $k < n$ random variables $X_1, \ldots, X_k$. Then for all values of $x_1, \ldots, x_k$ such that $f_0(x_1, \ldots, x_k) >$ 0, the conditional p.d.f. of $(X_{k+1}, \ldots, X_n)$ given that $X_1 = x_1, \ldots, X_k = x_k$ is defined

as follows:

$$g_{k+1\ldots n}(x_{k+1}, \ldots, x_n|x_1, \ldots, x_k) = \frac{f(x_1, x_2, \ldots, x_n)}{f_0(x_1, \ldots, x_k)}.$$

The definition above generalizes to arbitrary joint distributions as follows.

Definition 3.7.7 Conditional p.f., p.d.f., or p.f./p.d.f. Suppose that the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is divided into two subvectors $\boldsymbol{Y}$ and $\boldsymbol{Z}$, where $\boldsymbol{Y}$ is a k-dimensional random vector comprising k of the n random variables in $\boldsymbol{X}$, and $\boldsymbol{Z}$ is an $(n-k)$-dimensional random vector comprising the other $n-k$ random variables in $\boldsymbol{X}$. Suppose also that the n-dimensional joint p.f., p.d.f., or p.f./p.d.f. of $(\boldsymbol{Y}, \boldsymbol{Z})$ is f and that the marginal $(n-k)$-dimensional p.f., p.d.f., or p.f./p.d.f. of $\boldsymbol{Z}$ is f_2. Then for every given point $\boldsymbol{z} \in R^{n-k}$ such that $f_2(\boldsymbol{z}) > 0$, the conditional k-dimensional p.f., p.d.f., or p.f./p.d.f. g_1 of $\boldsymbol{Y}$ given $\boldsymbol{Z} = \boldsymbol{z}$ is defined as follows:

$$g_1(\boldsymbol{y}|\boldsymbol{z}) = \frac{f(\boldsymbol{y}, \boldsymbol{z})}{f_2(\boldsymbol{z})} \qquad \text{for } \boldsymbol{y} \in R^k. \tag{3.7.8}$$

Eq. (3.7.8) can be rewritten as

$$f(\boldsymbol{y}, \boldsymbol{z}) = g_1(\boldsymbol{y}|\boldsymbol{z}) f_2(\boldsymbol{z}), \tag{3.7.9}$$

which allows construction of the joint distribution from a conditional distribution and a marginal distribution. As in the bivariate case, it is safe to assume that $f(\boldsymbol{y}, \boldsymbol{z}) = 0$ whenever $f_2(\boldsymbol{z}) = 0$. Then Eq. (3.7.9) holds for all $\boldsymbol{y}$ and $\boldsymbol{z}$ even though $g_1(\boldsymbol{y}|\boldsymbol{z})$ is not uniquely defined.

Example 3.7.14 Service Times in a Queue. In Example 3.7.9, we calculated the marginal bivariate distribution of two service times $\boldsymbol{Z} = (X_1, X_2)$. We can now find the conditional three-dimensional p.d.f. of $\boldsymbol{Y} = (X_3, X_4, X_5)$ given $\boldsymbol{Z} = (x_1, x_2)$ for every pair (x_1, x_2) such that $x_1, x_2 > 0$:

$$\begin{aligned} g_1(x_3, x_4, x_5|x_1, x_2) &= \frac{f(x_1, \ldots, x_5)}{f_{12}(x_1, x_2)} \\ &= \left(\frac{240}{(2 + x_1 + \cdots + x_5)^6}\right)\left(\frac{4}{(2 + x_1 + x_2)^3}\right)^{-1} \\ &= \frac{60(2 + x_1 + x_2)^3}{(2 + x_1 + \cdots + x_5)^6}, \end{aligned} \tag{3.7.10}$$

for $x_3, x_4, x_5 > 0$, and 0 otherwise. The joint p.d.f. in (3.7.10) looks like a bunch of symbols, but it can be quite useful. Suppose that we observe $X_1 = 4$ and $X_2 = 6$. Then

$$g_1(x_3, x_4, x_5|4.6) = \begin{cases} \dfrac{103{,}680}{(12 + x_3 + x_4 + x_5)^6} & \text{for } x_3, x_4, x_5 > 0, \\ 0 & \text{otherwise.} \end{cases}$$

We can now calculate the conditional probability that $X_3 > 3$ given $X_1 = 4$, $X_2 = 6$:

$$\begin{aligned}\Pr(X_3 > 3|X_1 = 4,\ X_2 = 6) &= \int_3^\infty \int_0^\infty \int_0^\infty \frac{10{,}360}{(12 + x_3 + x_4 + x_5)^6} dx_5 dx_4 dx_3 \\ &= \int_3^\infty \int_0^\infty \frac{20{,}736}{(12 + x_3 + x_4)^5} dx_4 dx_3 \\ &= \int_3^\infty \frac{5184}{(12 + x_3)^4} dx_3 \\ &= \frac{1728}{15^3} = 0.512.\end{aligned}$$

Compare this to the calculation of $\Pr(X_3 > 3) = 0.4$ at the end of Example 3.7.9. After learning that the first two service times are a bit longer than three time units, we revise the probability that $X_3 > 3$ upward to reflect what we learned from the first two observations. If the first two service times had been small, the conditional probability that $X_3 > 3$ would have been smaller than 0.4. For example, $\Pr(X_3 > 3|X_1 = 1,\ X_2 = 1.5) = 0.216$. ◄

Example 3.7.15 Determining a Marginal Bivariate p.d.f. Suppose that Z is a random variable for which the p.d.f. f_0 is as follows:

$$f_0(z) = \begin{cases} 2e^{-2z} & \text{for } z > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.11}$$

Suppose, furthermore, that for every given value $Z = z > 0$ two other random variables X_1 and X_2 are independent and identically distributed and the conditional p.d.f. of each of these variables is as follows:

$$g(x|z) = \begin{cases} ze^{-zx} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.12}$$

We shall determine the marginal joint p.d.f. of (X_1, X_2).

Since X_1 and X_2 are i.i.d. for each given value of Z, their conditional joint p.d.f. when $Z = z > 0$ is

$$g_{12}(x_1, x_2|z) = \begin{cases} z^2 e^{-z(x_1+x_2)} & \text{for } x_1, x_2 > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The joint p.d.f. f of (Z, X_1, X_2) will be positive only at those points (z, x_1, x_2) such that $x_1, x_2, z > 0$. It now follows that, at every such point,

$$f(z, x_1, x_2) = f_0(z) g_{12}(x_1, x_2|z) = 2z^2 e^{-z(2+x_1+x_2)}.$$

For $x_1 > 0$ and $x_2 > 0$, the marginal joint p.d.f. $f_{12}(x_1, x_2)$ of X_1 and X_2 can be determined either using integration by parts or some special results that will arise in Sec. 5.7:

$$f_{12}(x_1, x_2) = \int_0^\infty f(z, x_1, x_2)\, dz = \frac{4}{(2 + x_1 + x_2)^3},$$

for $x_1, x_2 > 0$. The reader will note that this p.d.f. is the same as the marginal bivariate p.d.f. of (X_1, X_2) found in Eq. (3.7.5).

From this marginal bivariate p.d.f., we can evaluate probabilities involving X_1 and X_2, such as $\Pr(X_1 + X_2 < 4)$. We have

$$\Pr(X_1 + X_2 < 4) = \int_0^4 \int_0^{4-x_2} \frac{4}{(2 + x_1 + x_2)^3}\, dx_1\, dx_2 = \frac{4}{9}.$$ ◄

Example 3.7.16 Service Times in a Queue. We can think of the random variable Z in Example 3.7.15 as the rate at which customers are served in the queue of Example 3.7.5. With this interpretation, it is useful to find the conditional distribution of the rate Z after we observe some of the service times such as X_1 and X_2.

For every value of z, the conditional p.d.f. of Z given $X_1 = x_1$ and $X_2 = x_2$ is

$$\begin{aligned} g_0(z|x_1, x_2) &= \frac{f(z, x_1, x_2)}{f_{12}(x_1, x_2)} \\ &= \begin{cases} \frac{1}{2}(2 + x_1 + x_2)^3 z^2 e^{-z(2+x_1+x_2)} & \text{for } z > 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{3.7.13}$$

Finally, we shall evaluate $\Pr(Z \le 1|X_1 = 1,\ X_2 = 4)$. We have

$$\begin{aligned} \Pr(Z \le 1|X_1 = 1,\ X_2 = 4) &= \int_0^1 g_0(z|1, 4)\, dz \\ &= \int_0^1 171.5 z^2 e^{-7z}\, dz = 0.9704. \end{aligned}$$ ◀

Law of Total Probability and Bayes' Theorem Example 3.7.15 contains an example of the multivariate version of the law of total probability, while Example 3.7.16 contains an example of the multivariate version of Bayes' theorem. The proofs of the general versions are straightforward consequences of Definition 3.7.7.

Theorem 3.7.5 Multivariate Law of Total Probability and Bayes' Theorem. Assume the conditions and notation given in Definition 3.7.7. If $\boldsymbol{Z}$ has a continuous joint distribution, the marginal p.d.f. of $\boldsymbol{Y}$ is

$$f_1(\boldsymbol{y}) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-k} g_1(\boldsymbol{y}|\boldsymbol{z}) f_2(\boldsymbol{z})\, d\boldsymbol{z}, \tag{3.7.14}$$

and the conditional p.d.f. of $\boldsymbol{Z}$ given $\boldsymbol{Y} = \boldsymbol{y}$ is

$$g_2(\boldsymbol{z}|\boldsymbol{y}) = \frac{g_1(\boldsymbol{y}|\boldsymbol{z}) f_2(\boldsymbol{z})}{f_1(\boldsymbol{y})}. \tag{3.7.15}$$

If $\boldsymbol{Z}$ has a discrete joint distribution, then the multiple integral in (3.7.14) must be replaced by a multiple summation. If $\boldsymbol{Z}$ has a mixed joint distribution, the multiple integral must be replaced by integration over those coordinates with continuous distributions and summation over those coordinates with discrete distributions. ■

Conditionally Independent Random Variables In Examples 3.7.15 and 3.7.16, $\boldsymbol{Z}$ is the single random variable Z and $\boldsymbol{Y} = (X_1, X_2)$. These examples also illustrate the use of conditionally independent random variables. That is, X_1 and X_2 are conditionally independent given $Z = z$ for all $z > 0$. In Example 3.7.16, we said that Z was the rate at which customers were served. When this rate is unknown, it is a major source of uncertainty. Partitioning the sample space by the values of the rate Z and then conditioning on each value of Z removes a major source of uncertainty for part of the calculation.

In general, conditional independence for random variables is similar to conditional independence for events.

Definition 3.7.8 Conditionally Independent Random Variables. Let $\boldsymbol{Z}$ be a random vector with joint p.f., p.d.f., or p.f./p.d.f. $f_0(\boldsymbol{z})$. Several random variables $X_1, \ldots, X_n$ are *conditionally independent given* $\boldsymbol{Z}$ if, for all $\boldsymbol{z}$ such that $f_0(\boldsymbol{z}) > 0$, we have

$$g(\boldsymbol{x}|\boldsymbol{z}) = \prod_{i=1}^{n} g_i(x_i|\boldsymbol{z}),$$

where $g(\boldsymbol{x}|\boldsymbol{z})$ stands for the conditional multivariate p.f., p.d.f., or p.f./p.d.f. of $\boldsymbol{X}$ given $\boldsymbol{Z} = \boldsymbol{z}$ and $g_i(x_i|\boldsymbol{z})$ stands for the conditional univariate p.f. or p.d.f. of X_i given $\boldsymbol{Z} = \boldsymbol{z}$.

In Example 3.7.15, $g_i(x_i|z) = ze^{-zx_i}$ for $x_i > 0$ and $i = 1, 2$.

Example 3.7.17 A Clinical Trial. In Example 3.7.8, the joint p.f./p.d.f. given there was constructed by assuming that $X_1, \ldots, X_m$ were conditionally independent given $P = p$ each with the same conditional p.f., $g_i(x_i|p) = p^{x_i}(1-p)^{1-x_i}$ for $x_i \in \{0, 1\}$ and that P had the uniform distribution on the interval [0, 1]. These assumptions produce, in the notation of Definition 3.7.8,

$$g(\boldsymbol{x}|p) = \begin{cases} p^{x_1+\cdots+x_m}(1-p)^{40-x_1-\cdots-x_m} & \text{for all } x_i \in \{0, 1\} \text{ and } 0 \le p \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

for $0 \le p \le 1$. Combining this with the marginal p.d.f. of P, $f_2(p) = 1$ for $0 \le p \le 1$ and 0 otherwise, we get the joint p.f./p.d.f. given in Example 3.7.8. ◀

Conditional Versions of Past and Future Theorems We mentioned earlier that conditional distributions behave just like distributions. Hence, all theorems that we have proven and will prove in the future have conditional versions. For example, the law of total probability in Eq. (3.7.14) has the following version conditional on another random vector $\boldsymbol{W} = \boldsymbol{w}$:

$$f_1(\boldsymbol{y}|\boldsymbol{w}) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-k} g_1(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{w}) f_2(\boldsymbol{z}|\boldsymbol{w})\, d\boldsymbol{z}, \tag{3.7.16}$$

where $f_1(\boldsymbol{y}|\boldsymbol{w})$ stands for the conditional p.d.f., p.f., or p.f./p.d.f. of $\boldsymbol{Y}$ given $\boldsymbol{W} = \boldsymbol{w}$, $g_1(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{w})$ stands for the conditional p.d.f., p.f., or p.f./p.d.f. of $\boldsymbol{Y}$ given $(\boldsymbol{Z}, \boldsymbol{W}) = (\boldsymbol{z}, \boldsymbol{w})$, and $f_2(\boldsymbol{z}|\boldsymbol{w})$ stands for the conditional p.d.f. of $\boldsymbol{Z}$ given $\boldsymbol{W} = \boldsymbol{w}$. Using the same notation, the conditional version of Bayes' theorem is

$$g_2(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w}) = \frac{g_1(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{w}) f_2(\boldsymbol{z}|\boldsymbol{w})}{f_1(\boldsymbol{y}|\boldsymbol{w})}. \tag{3.7.17}$$

Example 3.7.18 Conditioning on Random Variables in Sequence. In Example 3.7.15, we found the conditional p.d.f. of Z given $(X_1, X_2) = (x_1, x_2)$. Suppose now that there are three more observations available, X_3, X_4, and X_5, and suppose that all of $X_1, \ldots, X_5$ are conditionally i.i.d. given $Z = z$ with p.d.f. $g(x|z)$. We shall use the conditional version of Bayes' theorem to compute the conditional p.d.f. of Z given $(X_1, \ldots, X_5) = (x_1, \ldots, x_5)$. First, we shall find the conditional p.d.f. $g_{345}(x_3, x_4, x_5|x_1, x_2, z)$ of $\boldsymbol{Y} = (X_3, X_4, X_5)$ given $Z = z$ and $\boldsymbol{W} = (X_1, X_2) = (x_1, x_2)$. We shall use the notation for p.d.f.'s in the discussion immediately preceding this example. Since $X_1, \ldots, X_5$ are conditionally i.i.d. given Z, we have that $g_1(\boldsymbol{y}|z, \boldsymbol{w})$ does not depend on $\boldsymbol{w}$. In fact,

$$g_1(\boldsymbol{y}|z, \boldsymbol{w}) = g(x_3|z)g(x_4|z)g(x_5|z) = z^3 e^{-z(x_3+x_4+x_5)},$$

for $x_3, x_4, x_5 > 0$. We also need the conditional p.d.f. of Z given $\boldsymbol{W} = \boldsymbol{w}$, which was calculated in Eq. (3.7.13), and we now denote it

$$f_2(z|\boldsymbol{w}) = \frac{1}{2}(2 + x_1 + x_2)^3 z^2 e^{-z(2+x_1+x_2)}.$$

Finally, we need the conditional p.d.f. of the last three observations given the first two. This was calculated in Example 3.7.14, and we now denote it

$$f_1(\boldsymbol{y}|\boldsymbol{w}) = \frac{60(2 + x_1 + x_2)^3}{(2 + x_1 + \cdots + x_5)^6}.$$

Now combine these using Bayes' theorem (3.7.17) to obtain

$$\begin{aligned} g_2(\boldsymbol{z}|\,\boldsymbol{y}, \boldsymbol{w}) &= \frac{z^3 e^{-z(x_3+x_4+x_5)} \frac{1}{2}(2 + x_1 + x_2)^3 z^2 e^{-z(2+x_1+x_2)}}{\dfrac{60(2 + x_1 + x_2)^3}{(2 + x_1 + \cdots + x_5)^6}} \\ &= \frac{1}{120}(2 + x_1 + \cdots + x_5)^6 z^5 e^{-z(2+x_1+\cdots+x_5)}, \end{aligned}$$

for $z > 0$. ◂

Note: Simple Rule for Creating Conditional Versions of Results. If you ever wish to determine the conditional version given $\boldsymbol{W} = \boldsymbol{w}$ of a result that you have proven, here is a simple method. Just add "conditional on $\boldsymbol{W} = \boldsymbol{w}$" to every probabilistic statement in the result. This includes all probabilities, c.d.f.'s, quantiles, names of distributions, p.d.f.'s, p.f.'s, and so on. It also includes all future probabilistic concepts that we introduce in later chapters (such as expected values and variances in Chapter 4).

Note: Independence is a Special Case of Conditional Independence. Let $X_1, \ldots,$ X_n be independent random variables, and let W be a constant random variable. That is, there is a constant c such that $\Pr(W = c) = 1$. Then $X_1, \ldots, X_n$ are also conditionally independent given $W = c$. The proof is straightforward and is left to the reader (Exercise 15). This result is not particularly interesting in its own right. Its value is the following: If we prove a result for conditionally independent random variables or conditionally i.i.d. random variables, then the same result will hold for independent random variables or i.i.d. random variables as the case may be.

Histograms

Example 3.7.19 Rate of Service. In Examples 3.7.5 and 3.7.6, we considered customers arriving at a queue and being served. Let Z stand for the rate at which customers were served, and we let $X_1, X_2, \ldots$ stand for the times that the successive customers requrired for service. Assume that $X_1, X_2, \ldots$ are conditionally i.i.d. given $Z = z$ with p.d.f.

$$g(x|z) = \begin{cases} ze^{-zx} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.18}$$

This is the same as (3.7.12) from Example 3.7.15. In that example, we modeled Z as a random variable with p.d.f. $f_0(z) = 2\exp(-2z)$ for $z > 0$. In this example, we shall assume that $X_1, \ldots, X_n$ will be observed for some large value n, and we want to think about what these observations tell us about Z. To be specific, suppose that we observe $n = 100$ service times. The first 10 times are listed here:

1.39, 0.61, 2.47, 3.35, 2.56, 3.60, 0.32, 1.43, 0.51, 0.94.

The smallest and largest observed service times from the entire sample are 0.004 and 9.60, respectively. It would be nice to have a graphical display of the entire sample of $n = 100$ service times without having to list them separately. ◀

The histogram, defined below, is a graphical display of a collection of numbers. It is particularly useful for displaying the observed values of a collection of random variables that have been modeled as conditionally i.i.d.

Definition 3.7.9 Histogram. Let $x_1, \ldots, x_n$ be a collection of numbers that all lie between two values $a < b$. That is, $a \le x_i \le b$ for all $i = 1, \ldots, n$. Choose some integer $k \ge 1$ and divide the interval $[a, b]$ into k equal-length subintervals of length $(b - a)/k$. For each subinterval, count how many of the numbers $x_1, \ldots, x_n$ are in the subinterval. Let c_i be the count for subinterval i for $i = 1, \ldots, k$. Choose a number $r > 0$. (Typically, $r = 1$ or $r = n$ or $r = n(b - a)/k$.) Draw a two-dimensional graph with the horizonal axis running from a to b. For each subinterval $i = 1, \ldots, k$ draw a rectangular bar of width $(b - a)/k$ and height equal to c_i/r over the midpoint of the ith interval. Such a graph is called a *histogram*.

The choice of the number r in the definition of histogram depends on what one wishes to be displayed on the vertical axis. The shape of the histogram is identical regardless of what value one chooses for r. With $r = 1$, the height of each bar is the raw count for each subinterval, and counts are displayed on the vertical axis. With $r = n$, the height of each bar is the proportion of the set of numbers in each subinterval, and the vertical axis displays proportions. With $r = n(b - a)/k$, the area of each bar is the proportion of the set of numbers in each subinterval.

Example 3.7.20 Rate of Service. The $n = 100$ observed service times in Example 3.7.19 all lie between 0 and 10. It is convenient, in this example, to draw a histogram with horizontal axis running from 0 to 10 and divided into 10 subintervals of length 1 each. Other choices are possible, but this one will do for illustration. Figure 3.22 contains the histogram of the 100 observed service times with $r = 100$. One sees that the numbers of observed service times in the subintervals decrease as the center of the subinterval increses. This matches the behavior of the conditional p.d.f. $g(x|z)$ of the service times as a function of x for fixed z. ◀

Histograms are useful as more than just graphical displays of large sets of numbers. After we see the law of large numbers (Theorem 6.2.4), we can show that the

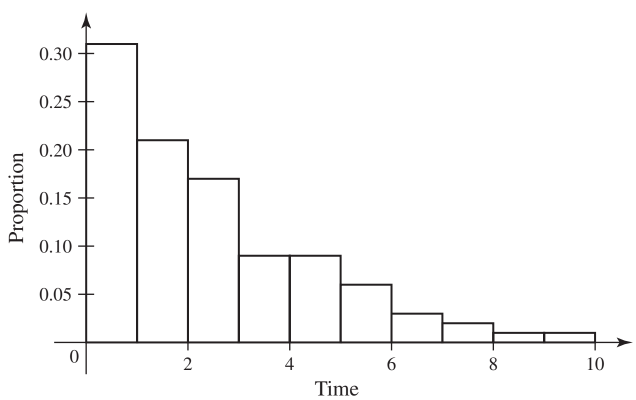


Figure 3.22 Histogram of service times for Example 3.7.20 with $a = 0$, $b = 10$, $k = 10$, and $r = 100$.

histogram of a large (conditionally) i.i.d. sample of continuous random variables is an approximation to the (conditional) p.d.f. of the random variables in the sample, so long as one uses the third choice of r, namely, $r = n(b - a)/k$.

Note: More General Histograms. Sometimes it is convenient to divide the range of the numbers to be plotted in a histogram into unequal-length subintervals. In such a case, one would typically let the height of each bar be c_i/r_i, where c_i is the raw count and r_i is proportional to the length of the ith subinterval. In this way, the area of each bar is still proportional to the count or proportion in each subinterval.

Summary

A finite collection of random variables is called a random vector. We have defined joint distributions for arbitrary random vectors. Every random vector has a joint c.d.f. Continuous random vectors have a joint p.d.f. Discrete random vectors have a joint p.f. Mixed distribution random vectors have a joint p.f./p.d.f. The coordinates of an n-dimensional random vector $\boldsymbol{X}$ are independent if the joint p.f., p.d.f., or p.f./p.d.f. $f(\boldsymbol{x})$ factors into $\prod_{i=1}^{n} f_i(x_i)$.

We can compute marginal distributions of subvectors of a random vector, and we can compute the conditional distribution of one subvector given the rest of the vector. We can construct a joint distribution for a random vector by piecing together a marginal distribution for part of the vector and a conditional distribution for the rest given the first part. There are versions of Bayes' theorem and the law of total probability for random vectors.

An n-dimensional random vector $\boldsymbol{X}$ has coordinates that are conditionally independent given $\boldsymbol{Z}$ if the conditional p.f., p.d.f., or p.f./p.d.f. $g(\boldsymbol{x}|\boldsymbol{z})$ of $\boldsymbol{X}$ given $\boldsymbol{Z} = \boldsymbol{z}$ factors into $\prod_{i=1}^{n} g_i(x_i|\boldsymbol{z})$. There are versions of Bayes' theorem, the law of total probability, and all future theorems about random variables and random vectors conditional on an arbitrary additional random vector.

Exercises

1. Suppose that three random variables X_1, X_2, and X_3 have a continuous joint distribution with the following joint p.d.f.: $f(x_1, x_2, x_3) =$

$$\begin{cases} c(x_1 + 2x_2 + 3x_3) & \text{for } 0 \le x_i \le 1 \ (i = 1, 2, 3), \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the value of the constant c;
(b) the marginal joint p.d.f. of X_1 and X_3; and
(c) $\Pr\left(X_3 < \frac{1}{2}\middle| X_1 = \frac{1}{4},\ X_2 = \frac{3}{4}\right)$.

2. Suppose that three random variables X_1, X_2, and X_3 have a mixed joint distribution with p.f./p.d.f.:

$$f(x_1, x_2, x_3) = \begin{cases} c x_1^{1+x_2+x_3}(1 - x_1)^{3-x_2-x_3} & \text{if } 0 < x_1 < 1 \\ & \quad \text{and } x_2, x_3 \in \{0, 1\}, \\ 0 & \text{otherwise.} \end{cases}$$

(Notice that X_1 has a continuous distribution and X_2 and X_3 have discrete distributions.) Determine **(a)** the value of the constant c; **(b)** the marginal joint p.f. of X_2 and X_3; and **(c)** the conditional p.d.f. of X_1 given $X_2 = 1$ and $X_3 = 1$.

3. Suppose that three random variables X_1, X_2, and X_3 have a continuous joint distribution with the following joint p.d.f.: $f(x_1, x_2, x_3) =$

$$\begin{cases} ce^{-(x_1+2x_2+3x_3)} & \text{for } x_i > 0 \ (i = 1, 2, 3), \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the value of the constant c; **(b)** the marginal joint p.d.f. of X_1 and X_3; and **(c)** $\Pr(X_1 < 1|X_2 = 2,\ X_3 = 1)$.

4. Suppose that a point (X_1, X_2, X_3) is chosen at random, that is, in accordance with the uniform p.d.f., from the following set S:

$$S = \{(x_1, x_2, x_3) : 0 \le x_i \le 1 \text{ for } i = 1, 2, 3\}.$$

Determine:

a. $\Pr\left[\left(X_1-\frac{1}{2}\right)^2+\left(X_2-\frac{1}{2}\right)^2+\left(X_3-\frac{1}{2}\right)^2\le\frac{1}{4}\right]$

b. $\Pr(X_1^2+X_2^2+X_3^2\le 1)$

5. Suppose that an electronic system contains n components that function independently of each other and that the probability that component i will function properly is p_i $(i=1,\ldots,n)$. It is said that the components are connected *in series* if a necessary and sufficient condition for the system to function properly is that all n components function properly. It is said that the components are connected *in parallel* if a necessary and sufficient condition for the system to function properly is that at least one of the n components functions properly. The probability that the system will function properly is called the *reliability* of the system. Determine the reliability of the system, **(a)** assuming that the components are connected in series, and **(b)** assuming that the components are connected in parallel.

6. Suppose that the n random variables $X_1\ldots,X_n$ form a random sample from a discrete distribution for which the p.f. is f. Determine the value of $\Pr(X_1=X_2=\cdots=X_n)$.

7. Suppose that the n random variables $X_1,\ldots,X_n$ form a random sample from a continuous distribution for which the p.d.f. is f. Determine the probability that at least k of these n random variables will lie in a specified interval $a\le x\le b$.

8. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x)=\begin{cases}\frac{1}{n!}x^n e^{-x} & \text{for } x>0\\ 0 & \text{otherwise.}\end{cases}$$

Suppose also that for any given value $X=x$ $(x>0)$, the n random variables $Y_1,\ldots,Y_n$ are i.i.d. and the conditional p.d.f. g of each of them is as follows:

$$g(y|x)=\begin{cases}\frac{1}{x} & \text{for } 0<y<x,\\ 0 & \text{otherwise.}\end{cases}$$

Determine **(a)** the marginal joint p.d.f. of $Y_1,\ldots,Y_n$ and **(b)** the conditional p.d.f. of X for any given values of $Y_1,\ldots,Y_n$.

9. Let X be a random variable with a continuous distribution. Let $X_1=X_2=X$.

a. Prove that both X_1 and X_2 have a continuous distribution.

b. Prove that $\boldsymbol{X}=(X_1,X_2)$ does not have a continuous joint distribution.

10. Return to the situation described in Example 3.7.18. Let $\boldsymbol{X}=(X_1,\ldots,X_5)$ and compute the conditional p.d.f. of Z given $\boldsymbol{X}=\boldsymbol{x}$ directly in one step, as if all of $\boldsymbol{X}$ were observed at the same time.

11. Suppose that $X_1,\ldots,X_n$ are independent. Let $k<n$ and let $i_1,\ldots,i_k$ be distinct integers between 1 and n. Prove that $X_{i_1},\ldots,X_{i_k}$ are independent.

12. Let $\boldsymbol{X}$ be a random vector that is split into three parts, $\boldsymbol{X}=(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{W})$. Suppose that $\boldsymbol{X}$ has a continuous joint distribution with p.d.f. $f(\boldsymbol{y},\boldsymbol{z},\boldsymbol{w})$. Let $g_1(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{w})$ be the conditional p.d.f. of $(\boldsymbol{Y},\boldsymbol{Z})$ given $\boldsymbol{W}=\boldsymbol{w}$, and let $g_2(\boldsymbol{y}|\boldsymbol{w})$ be the conditional p.d.f. of $\boldsymbol{Y}$ given $\boldsymbol{W}=\boldsymbol{w}$. Prove that $g_2(\boldsymbol{y}|\boldsymbol{w})=\int g_1(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{w})\,d\boldsymbol{z}$.

13. Let X_1,X_2,X_3 be conditionally independent given $Z=z$ for all z with the conditional p.d.f. $g(x|z)$ in Eq. (3.7.12). Also, let the marginal p.d.f. of Z be f_0 in Eq. (3.7.11). Prove that the conditional p.d.f. of X_3 given $(X_1,X_2)=(x_1,x_2)$ is $\int_0^\infty g(x_3|z)g_0(z|x_1,x_2)\,dz$, where g_0 is defined in Eq. (3.7.13). (You can prove this even if you cannot compute the integral in closed form.)

14. Consider the situation described in Example 3.7.14. Suppose that $X_1=5$ and $X_2=7$ are observed.

a. Compute the conditional p.d.f. of X_3 given $(X_1,X_2)=(5,7)$. (You may use the result stated in Exercise 12.)

b. Find the conditional probability that $X_3>3$ given $(X_1,X_2)=(5,7)$ and compare it to the value of $\Pr(X_3>3)$ found in Example 3.7.9. Can you suggest a reason why the conditional probability should be higher than the marginal probability?

15. Let $X_1,\ldots,X_n$ be independent random variables, and let W be a random variable such that $\Pr(W=c)=1$ for some constant c. Prove that $X_1,\ldots,X_n$ are conditionally independent given $W=c$.

3.8 Functions of a Random Variable

Often we find that after we compute the distribution of a random variable X, we really want the distribution of some function of X. For example, if X is the rate at which customers are served in a queue, then $1/X$ is the average waiting time. If we have the distribution of X, we should be able to determine the distribution of $1/X$ or of any other function of X. How to do that is the subject of this section.

Random Variable with a Discrete Distribution

Example 3.8.1 Distance from the Middle. Let X have the uniform distribution on the integers $1, 2, \ldots, 9$. Suppose that we are interested in how far X is from the middle of the distribution, namely, 5. We could define $Y = |X - 5|$ and compute probabilities such as $\Pr(Y = 1) = \Pr(X \in \{4, 6\}) = 2/9$. ◄

Example 3.8.1 illustrates the general procedure for finding the distribution of a function of a discrete random variable. The general result is straightforward.

Theorem 3.8.1 Function of a Discrete Random Variable. Let X have a discrete distribution with p.f. f, and let $Y = r(X)$ for some function of r defined on the set of possible values of X. For each possible value y of Y, the p.f. g of Y is

$$g(y) = \Pr(Y = y) = \Pr[r(X) = y] = \sum_{x: r(x) = y} f(x). \quad \blacksquare$$

Example 3.8.2 Distance from the Middle. The possible values of Y in Example 3.8.1 are 0, 1, 2, 3, and 4. We see that $Y = 0$ if and only if $X = 5$, so $g(0) = f(5) = 1/9$. For all other values of Y, there are two values of X that give that value of Y. For example, $\{Y = 4\} = \{X = 1\} \cup \{X = 9\}$. So, $g(y) = 2/9$ for $y = 1, 2, 3, 4$. ◄

Random Variable with a Continuous Distribution

If a random variable X has a continuous distribution, then the procedure for deriving the probability distribution of a function of X differs from that given for a discrete distribution. One way to proceed is by direct calculation as in Example 3.8.3.

Example 3.8.3 Average Waiting Time. Let Z be the rate at which customers are served in a queue, and suppose that Z has a continuous c.d.f. F. The average waiting time is $Y = 1/Z$. If we want to find the c.d.f. G of Y, we can write

$$G(y) = \Pr(Y \le y) = \Pr\left(\frac{1}{Z} \le y\right) = \Pr\left(Z \ge \frac{1}{y}\right) = \Pr\left(Z > \frac{1}{y}\right) = 1 - F\left(\frac{1}{y}\right),$$

where the fourth equality follows from the fact that Z has a continuous distribution so that $\Pr(Z = 1/y) = 0$. ◄

In general, suppose that the p.d.f. of X is f and that another random variable is defined as $Y = r(X)$. For each real number y, the c.d.f. $G(y)$ of Y can be derived as follows:

$$\begin{aligned} G(y) &= \Pr(Y \le y) = \Pr[r(X) \le y] \\ &= \int_{\{x: r(x) \le y\}} f(x)\, dx. \end{aligned}$$

If the random variable Y also has a continuous distribution, its p.d.f. g can be obtained from the relation

$$g(y) = \frac{dG(y)}{dy}.$$

This relation is satisfied at every point y at which G is differentiable.

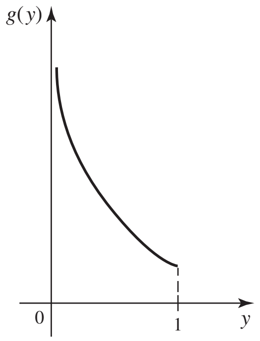


Figure 3.23 The p.d.f. of $Y = X^2$ in Example 3.8.4.

Example 3.8.4 Deriving the p.d.f. of X^2 when X Has a Uniform Distribution. Suppose that X has the uniform distribution on the interval $[-1, 1]$, so

$$f(x) = \begin{cases} 1/2 & \text{for } -1 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We shall determine the p.d.f. of the random variable $Y = X^2$.

Since $Y = X^2$, then Y must belong to the interval $0 \le Y \le 1$. Thus, for each value of Y such that $0 \le y \le 1$, the c.d.f. $G(y)$ of Y is

$$\begin{aligned} G(y) &= \Pr(Y \le y) = \Pr(X^2 \le y) \\ &= \Pr(-y^{1/2} \le X \le y^{1/2}) \\ &= \int_{-y^{1/2}}^{y^{1/2}} f(x)\,dx = y^{1/2}. \end{aligned}$$

For $0 < y < 1$, it follows that the p.d.f. $g(y)$ of Y is

$$g(y) = \frac{dG(y)}{dy} = \frac{1}{2y^{1/2}}.$$

This p.d.f. of Y is sketched in Fig. 3.23. It should be noted that although Y is simply the square of a random variable with a uniform distribution, the p.d.f. of Y is unbounded in the neighborhood of $y = 0$. ◂

Linear functions are very useful transformations, and the p.d.f. of a linear function of a continuous random variable is easy to derive. The proof of the following result is left to the reader in Exercise 5.

Theorem 3.8.2 Linear Function. Suppose that X is a random variable for which the p.d.f. is f and that $Y = aX + b$ $(a \ne 0)$. Then the p.d.f. of Y is

$$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right) \quad \text{for } -\infty < y < \infty, \tag{3.8.1}$$

and 0 otherwise. ■

The Probability Integral Transformation

Example 3.8.5 Let X be a continuous random variable with p.d.f. $f(x) = \exp(-x)$ for $x > 0$ and 0 otherwise. The c.d.f. of X is $F(x) = 1 - \exp(-x)$ for $x > 0$ and 0 otherwise. If we let

F be the function r in the earlier results of this section, we can find the distribution of $Y = F(X)$. The c.d.f. or Y is, for $0 < y < 1$,

$$\begin{aligned} G(y) &= \Pr(Y \le y) = \Pr(1 - \exp(-X) \le y) = \Pr(X \le -\log(1 - y)) \\ &= F(-\log(1 - y)) = 1 - \exp(-[-\log(1 - y)]) = y, \end{aligned}$$

which is the c.d.f. of the uniform distribution on the interval [0, 1]. It follows that Y has the uniform distribution on the interval [0, 1]. ◂

The result in Example 3.8.5 is quite general.

Theorem 3.8.3 Probability Integral Transformation. Let X have a continuous c.d.f. F, and let $Y = F(X)$. (This transformation from X to Y is called the *probability integral transformation*.) The distribution of Y is the uniform distribution on the interval [0, 1].

Proof First, because F is the c.d.f. of a random variable, then $0 \le F(x) \le 1$ for $-\infty < x < \infty$. Therefore, $\Pr(Y < 0) = \Pr(Y > 1) = 0$. Since F is continuous, the set of x such that $F(x) = y$ is a nonempty closed and bounded interval $[x_0, x_1]$ for each y in the interval (0, 1). Let $F^{-1}(y)$ denote the lower endpoint x_0 of this interval, which was called the y quantile of F in Definition 3.3.2. In this way, $Y \le y$ if and only if $X \le x_1$. Let G denote the c.d.f. of Y. Then

$$G(y) = \Pr(Y \le y) = \Pr(X \le x_1) = F(x_1) = y.$$

Hence, $G(y) = y$ for $0 < y < 1$. Because this function is the c.d.f. of the uniform distribution on the interval [0, 1], this uniform distribution is the distribution of Y. ■

Because $\Pr(X = F^{-1}(Y)) = 1$ in the proof of Theorem 3.8.3, we have the following corollary.

Corollary 3.8.1 Let Y have the uniform distribution on the interval [0, 1], and let F be a continuous c.d.f. with quantile function F^{-1}. Then $X = F^{-1}(Y)$ has c.d.f. F. ■

Theorem 3.8.3 and its corollary give us a method for transforming an arbitrary continuous random variable X into another random variable Z with any desired continuous distribution. To be specific, let X have a continuous c.d.f. F, and let G be another continuous c.d.f. Then $Y = F(X)$ has the uniform distribution on the interval [0, 1] according to Theorem 3.8.3, and $Z = G^{-1}(Y)$ has the c.d.f. G according to Corollary 3.8.1. Combining these, we see that $Z = G^{-1}[F(X)]$ has c.d.f. G.

Simulation

Pseudo-Random Numbers Most computer packages that do statistical analyses also produce what are called *pseudo-random numbers*. These numbers appear to have some of the properties that a random sample would have, even though they are generated by deterministic algorithms. The most fundamental of these programs are the ones that generate pseudo-random numbers that appear to have the uniform distribution on the interval [0, 1]. We shall refer to such functions as *uniform pseudo-random number generators*. The important features that a uniform pseudo-random number generator must have are the following.

The numbers that it produces need to be spread somewhat uniformly over the interval [0, 1], and they need to appear to be observed values of independent random

variables. This last feature is very complicated to word precisely. An example of a sequence that does *not* appear to be observations of independent random variables would be one that was perfectly evenly spaced. Another example would be one with the following behavior: Suppose that we look at the sequence $X_1, X_2, \ldots$ one at a time, and every time we find an $X_i > 0.5$, we write down the next number X_{i+1}. If the subsequence of numbers that we write down is not spread approximately uniformly over the interval [0, 1], then the original sequence does not look like observations of independent random variables with the uniform distribution on the interval [0, 1]. The reason is that the conditional distribution of X_{i+1} given that $X_i > 0.5$ is supposed to be uniform over the interval [0, 1], according to independence.

Generating Pseudo-Random Numbers Having a Specified Distribution A uniform pseudo-random number generator can be used to generate values of a random variable Y having any specified continuous c.d.f. G. If a random variable X has the uniform distribution on the interval [0, 1] and if the quantile function G^{-1} is defined as before, then it follows from Corollary 3.8.1 that the c.d.f. of the random variable $Y = G^{-1}(X)$ will be G. Hence, if a value of X is produced by a uniform pseudo-random number generator, then the corresponding value of Y will have the desired property. If n independent values $X_1, \ldots, X_n$ are produced by the generator, then the corresponding values $Y_1, \ldots, Y_n$ will appear to form a random sample of size n from the distribution with the c.d.f. G.

Example 3.8.6 Generating Independent Values from a Specified p.d.f. Suppose that a uniform pseudo-random number generator is to be used to generate three independent values from the distribution for which the p.d.f. g is as follows:

$$g(y) = \begin{cases} \frac{1}{2}(2 - y) & \text{for } 0 < y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

For $0 < y < 2$, the c.d.f. G of the given distribution is

$$G(y) = y - \frac{y^2}{4}.$$

Also, for $0 < x < 1$, the inverse function $y = G^{-1}(x)$ can be found by solving the equation $x = G(y)$ for y. The result is

$$y = G^{-1}(x) = 2[1 - (1 - x)^{1/2}]. \tag{3.8.2}$$

The next step is to generate three uniform pseudo-random numbers x_1, x_2, and x_3 using the generator. Suppose that the three generated values are

$$x_1 = 0.4125, \quad x_2 = 0.0894, \quad x_3 = 0.8302.$$

When these values of x_1, x_2, and x_3 are substituted successively into Eq. (3.8.2), the values of y that are obtained are $y_1 = 0.47$, $y_2 = 0.09$, and $y_3 = 1.18$. These are then treated as the observed values of three independent random variables with the distribution for which the p.d.f. is g. ◀

If G is a general c.d.f., there is a method similar to Corollary 3.8.1 that can be used to transform a uniform random variable into a random variable with c.d.f. G. See Exercise 12 in this section. There are other computer methods for generating values from certain specified distributions that are faster and more accurate than using the quantile function. These topics are discussed in the books by Kennedy and

Gentle (1980) and Rubinstein (1981). Chapter 12 of this text contains techniques and examples that show how simulation can be used to solve statistical problems.

General Function In general, if X has a continuous distribution and if $Y = r(X)$, then it is not necessarily true that Y will also have a continuous distribution. For example, suppose that $r(x) = c$, where c is a constant, for all values of x in some interval $a \le x \le b$, and that $\Pr(a \le X \le b) > 0$. Then $\Pr(Y = c) > 0$. Since the distribution of Y assigns positive probability to the value c, this distribution cannot be continuous. In order to derive the distribution of Y in a case like this, the c.d.f. of Y must be derived by applying methods like those described above. For certain functions r, however, the distribution of Y will be continuous; and it will then be possible to derive the p.d.f. of Y directly without first deriving its c.d.f. We shall develop this case in detail at the end of this section.

Direct Derivation of the p.d.f. When *r* is One-to-One and Differentiable

Example 3.8.7 Average Waiting Time. Consider Example 3.8.3 again. The p.d.f. g of Y can be computed from $G(y) = 1 - F(1/y)$ because F and $1/y$ both have derivatives at enough places. We apply the chain rule for differentiation to obtain

$$g(y) = \frac{dG(y)}{dy} = - \left.\frac{dF(x)}{dx}\right|_{x=1/y} \left(-\frac{1}{y^2}\right) = f\left(\frac{1}{y}\right)\frac{1}{y^2},$$

except at $y = 0$ and at those values of y such that $F(x)$ is not differentiable at $x = 1/y$. ◄

Differentiable One-To-One Functions The method used in Example 3.8.7 generalizes to very arbitrary differentiable one-to-one functions. Before stating the general result, we should recall some properties of differentiable one-to-one functions from calculus. Let r be a differentiable one-to-one function on the open interval (a, b). Then r is either strictly increasing or strictly decreasing. Because r is also continuous, it will map the interval (a, b) to another open interval (α, β), called the *image of* (a, b) *under* r. That is, for each $x \in (a, b)$, $r(x) \in (\alpha, \beta)$, and for each $y \in (\alpha, \beta)$ there is $x \in (a, b)$ such that $y = r(x)$ and this y is unique because r is one-to-one. So the inverse s of r will exist on the interval (α, β), meaning that for $x \in (a, b)$ and $y \in (\alpha, \beta)$ we have $r(x) = y$ if and only if $s(y) = x$. The derivative of s will exist (possibly infinite), and it is related to the derivative of r by

$$\frac{ds(y)}{dy} = \left(\left.\frac{dr(x)}{dx}\right|_{x=s(y)}\right)^{-1}.$$

Theorem 3.8.4 Let X be a random variable for which the p.d.f. is f and for which $\Pr(a < X < b) = 1$. (Here, a and/or b can be either finite or infinite.) Let $Y = r(X)$, and suppose that $r(x)$ is differentiable and one-to-one for $a < x < b$. Let (α, β) be the image of the interval (a, b) under the function r. Let $s(y)$ be the inverse function of $r(x)$ for $\alpha < y < \beta$. Then the p.d.f. g of Y is

$$g(y) = \begin{cases} f[s(y)]\left|\dfrac{ds(y)}{dy}\right| & \text{for } \alpha < y < \beta, \\ 0 & \text{otherwise.} \end{cases} \tag{3.8.3}$$

Proof If r is increasing, then s is increasing, and for each $y \in (\alpha, \beta)$,

$$G(y) = \Pr(Y \le y) = \Pr[r(X) \le y] = \Pr[X \le s(y)] = F[s(y)].$$

It follows that G is differentiable at all y where both s is differentiable and where $F(x)$ is differentiable at $x = s(y)$. Using the chain rule for differentiation, it follows that the p.d.f. $g(y)$ for $\alpha < y < \beta$ will be

$$g(y) = \frac{dG(y)}{dy} = \frac{dF[s(y)]}{dy} = f[s(y)]\,\frac{ds(y)}{dy}. \tag{3.8.4}$$

Because s is increasing, $ds(y)/dy$ is positive; hence, it equals $|ds(y)/dy|$ and Eq. (3.8.4) implies Eq. (3.8.3). Similarly, if r is decreasing, then s is decreasing, and for each $y \in (\alpha, \beta)$,

$$G(y) = \Pr[r(X) \le y] = \Pr[X \ge s(y)] = 1 - F[s(y)].$$

Using the chain rule again, we differentiate G to get the p.d.f. of Y

$$g(y) = \frac{dG(y)}{dy} = -f[s(y)]\,\frac{ds(y)}{dy}. \tag{3.8.5}$$

Since s is strictly decreasing, $ds(y)/dy$ is negative so that $-ds(y)/dy$ equals $|ds(y)/dy|$. It follows that Eq. (3.8.5) implies Eq. (3.8.3). ∎

Example 3.8.8

Microbial Growth. A popular model for populations of microscopic organisms in large environments is exponential growth. At time 0, suppose that v organisms are introduced into a large tank of water, and let X be the rate of growth. After time t, we would predict a population size of ve^{Xt}. Assume that X is unknown but has a continuous distribution with p.d.f.

$$f(x) = \begin{cases} 3(1-x)^2 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

We are interested in the distribution of $Y = ve^{Xt}$ for known values of v and t. For concreteness, let $v = 10$ and $t = 5$, so that $r(x) = 10e^{5x}$.

In this example, $\Pr(0 < X < 1) = 1$ and r is a continuous and strictly increasing function of x for $0 < x < 1$. As x varies over the interval $(0, 1)$, it is found that $y = r(x)$ varies over the interval $(10, 10e^5)$. Furthermore, for $10 < y < 10e^5$, the inverse function is $s(y) = \log(y/10)/5$. Hence, for $10 < y < 10e^5$,

$$\frac{ds(y)}{dy} = \frac{1}{5y}.$$

It follows from Eq. (3.8.3) that $g(y)$ will be

$$g(y) = \begin{cases} \dfrac{3(1 - \log(y/10)/5)^2}{5y} & \text{for } 10 < y < 10e^5, \\ 0 & \text{otherwise.} \end{cases}$$ ◂

Summary

We learned several methods for determining the distribution of a function of a random variable. For a random variable X with a continuous distribution having p.d.f. f, if r is strictly increasing or strictly decreasing with differentiable inverse s (i.e., $s(r(x)) = x$ and s is differentiable), then the p.d.f. of $Y = r(X)$ is $g(y) =$

$f(s(y))|ds(y)/dy|$. A special transformation allows us to transform a random variable X with the uniform distribution on the interval [0, 1] into a random variable Y with an arbitrary continuous c.d.f. G by $Y = G^{-1}(X)$. This method can be used in conjunction with a uniform pseudo-random number generator to generate random variables with arbitrary continuous distributions.

Exercises

1. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} 3x^2 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Also, suppose that $Y = 1 - X^2$. Determine the p.d.f. of Y.

2. Suppose that a random variable X can have each of the seven values $-3, -2, -1, 0, 1, 2, 3$ with equal probability. Determine the p.f. of $Y = X^2 - X$.

3. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{1}{2}x & \text{for } 0 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Also, suppose that $Y = X(2 - X)$. Determine the c.d.f. and the p.d.f. of Y.

4. Suppose that the p.d.f. of X is as given in Exercise 3. Determine the p.d.f. of $Y = 4 - X^3$.

5. Prove Theorem 3.8.2. (*Hint:* Either apply Theorem 3.8.4 or first compute the c.d.f. seperately for $a > 0$ and $a < 0$.)

6. Suppose that the p.d.f. of X is as given in Exercise 3. Determine the p.d.f. of $Y = 3X + 2$.

7. Suppose that a random variable X has the uniform distribution on the interval [0, 1]. Determine the p.d.f. of **(a)** X^2, **(b)** $-X^3$, and **(c)** $X^{1/2}$.

8. Suppose that the p.d.f. of X is as follows:

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0, \\ 0 & \text{for } x \le 0. \end{cases}$$

Determine the p.d.f. of $Y = X^{1/2}$.

9. Suppose that X has the uniform distribution on the interval [0, 1]. Construct a random variable $Y = r(X)$ for which the p.d.f. will be

$$g(y) = \begin{cases} \frac{3}{8}y^2 & \text{for } 0 < y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

10. Let X be a random variable for which the p.d.f f is as given in Exercise 3. Construct a random variable $Y = r(X)$ for which the p.d.f. g is as given in Exercise 9.

11. Explain how to use a uniform pseudo-random number generator to generate four independent values from a distribution for which the p.d.f. is

$$g(y) = \begin{cases} \frac{1}{2}(2y + 1) & \text{for } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

12. Let F be an arbitrary c.d.f. (not necessarily discrete, not necessarily continuous, not necessarily either). Let F^{-1} be the quantile function from Definition 3.3.2. Let X have the uniform distribution on the interval [0, 1]. Define $Y = F^{-1}(X)$. Prove that the c.d.f. of Y is F. *Hint:* Compute $\Pr(Y \le y)$ in two cases. First, do the case in which y is the unique value of x such that $F(x) = F(y)$. Second, do the case in which there is an entire interval of x values such that $F(x) = F(y)$.

13. Let Z be the rate at which customers are served in a queue. Assume that Z has the p.d.f.

$$f(z) = \begin{cases} 2e^{-2z} & \text{for } z > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Find the p.d.f. of the average waiting time $T = 1/Z$.

14. Let X have the uniform distribution on the interval $[a, b]$, and let $c > 0$. Prove that $cX + d$ has the uniform distribution on the interval $[ca + d, cb + d]$.

15. Most of the calculation in Example 3.8.4 is quite general. Suppose that X has a continuous distribution with p.d.f. f. Let $Y = X^2$, and show that the p.d.f. of Y is

$$g(y) = \frac{1}{2y^{1/2}}[f(y^{1/2}) + f(-y^{1/2})].$$

16. In Example 3.8.4, the p.d.f. of $Y = X^2$ is much larger for values of y near 0 than for values of y near 1 despite the fact that the p.d.f. of X is flat. Give an intuitive reason why this occurs in this example.

17. An insurance agent sells a policy which has a \$100 deductible and a \$5000 cap. This means that when the policy holder files a claim, the policy holder must pay the first

\$100. After the first \$100, the insurance company pays the rest of the claim up to a maximum payment of \$5000. Any excess must be paid by the policy holder. Suppose that the dollar amount X of a claim has a continuous distribution with p.d.f. $f(x) = 1/(1+x)^2$ for $x > 0$ and 0 otherwise. Let Y be the amount that the insurance company has to pay on the claim.

a. Write Y as a function of X, i.e., $Y = r(X)$.

b. Find the c.d.f. of Y.

c. Explain why Y has neither a continuous nor a discrete distribution.

3.9 Functions of Two or More Random Variables

When we observe data consisting of the values of several random variables, we need to summarize the observed values in order to be able to focus on the information in the data. Summarizing consists of constructing one or a few functions of the random variables that capture the bulk of the information. In this section, we describe the techniques needed to determine the distribution of a function of two or more random variables.

Random Variables with a Discrete Joint Distribution

Example 3.9.1

Bull Market. Three different investment firms are trying to advertise their mutual funds by showing how many perform better than a recognized standard. Each company has 10 funds, so there are 30 in total. Suppose that the first 10 funds belong to the first firm, the next 10 to the second firm, and the last 10 to the third firm. Let $X_i = 1$ if fund i performs better than the standard and $X_i = 0$ otherwise, for $i = 1, \dots, 30$. Then, we are interested in the three functions

$$
\begin{aligned}
Y_1 &= X_1 + \cdots + X_{10},\\
Y_2 &= X_{11} + \cdots + X_{20},\\
Y_3 &= X_{21} + \cdots + X_{30}.
\end{aligned}
$$

We would like to be able to determine the joint distribution of Y_1, Y_2, and Y_3 from the joint distribution of $X_1, \dots, X_{30}$. ◄

The general method for solving problems like those of Example 3.9.1 is a straightforward extension of Theorem 3.8.1.

Theorem 3.9.1

Functions of Discrete Random Variables. Suppose that n random variables $X_1, \dots, X_n$ have a discrete joint distribution for which the joint p.f. is f, and that m functions $Y_1, \dots, Y_m$ of these n random variables are defined as follows:

$$
\begin{aligned}
Y_1 &= r_1(X_1, \dots, X_n),\\
Y_2 &= r_2(X_1, \dots, X_n),\\
&\;\;\vdots\\
Y_m &= r_m(X_1, \dots, X_n).
\end{aligned}
$$

For given values $y_1, \ldots, y_m$ of the m random variables $Y_1, \ldots, Y_m$, let A denote the set of all points $(x_1, \ldots, x_n)$ such that

$$\begin{aligned} r_1(x_1, \ldots, x_n) &= y_1, \\ r_2(x_1, \ldots, x_n) &= y_2, \\ &\vdots \\ r_m(x_1, \ldots, x_n) &= y_m. \end{aligned}$$

Then the value of the joint p.f. g of $Y_1, \ldots, Y_m$ is specified at the point $(y_1, \ldots, y_m)$ by the relation

$$g(y_1, \ldots, y_m) = \sum_{(x_1, \ldots, x_n) \in A} f(x_1, \ldots, x_n).$$

■

Example 3.9.2 Bull Market. Recall the situation in Example 3.9.1. Suppose that we want the joint p.f. g of (Y_1, Y_2, Y_3) at the point $(3, 5, 8)$. That is, we want $g(3, 5, 8) = \Pr(Y_1 = 3, Y_2 = 5, Y_3 = 8)$. The set A as defined in Theorem 3.9.1 is

$$A = \{(x_1, \ldots, x_{30}) : x_1 + \cdots + x_{10} = 3, x_{11} + \cdots + x_{20} = 5, x_{21} + \cdots + x_{30} = 8\}.$$

Two of the points in the set A are

$$(1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0),$$
$$(1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1).$$

A counting argument like those developed in Sec. 1.8 can be used to discover that there are

$$\binom{10}{3}\binom{10}{5}\binom{10}{8} = 1{,}360{,}800$$

points in A. Unless the joint distribution of $X_1, \ldots, X_{30}$ has some simple structure, it will be extremely tedious to compute $g(3, 5, 8)$ as well as most other values of g. For example, if all of the 2^{30} possible values of the vector $(X_1, \ldots, X_{30})$ are equally likely, then

$$g(3, 5, 8) = \frac{1{,}360{,}800}{2^{30}} = 1.27 \times 10^{-3}.$$

◀

The next result gives an important example of a function of discrete random variables.

Theorem 3.9.2 Binomial and Bernoulli Distributions. Assume that $X_1, \ldots, X_n$ are i.i.d. random variables having the Bernoulli distribution with parameter p. Let $Y = X_1 + \cdots + X_n$. Then Y has the binomial distribution with parameters n and p.

Proof It is clear that $Y = y$ if and only if exactly y of $X_1, \ldots, X_n$ equal 1 and the other $n - y$ equal 0. There are $\binom{n}{y}$ distinct possible values for the vector $(X_1, \ldots, X_n)$ that have y ones and $n - y$ zeros. Each such vector has probability $p^y(1-p)^{n-y}$ of being observed; hence the probability that $Y = y$ is the sum of the probabilities of those vectors, namely, $\binom{n}{y}p^y(1-p)^{n-y}$ for $y = 0, \ldots, n$. From Definition 3.1.7, we see that Y has the binomial distribution with parameters n and p. ■

Example 3.9.3 Sampling Parts. Suppose that two machines are producing parts. For $i = 1, 2$, the probability is p_i that machine i will produce a defective part, and we shall assume that all parts from both machines are independent. Assume that the first n_1 parts are produced by machine 1 and that the last n_2 parts are produced by machine 2,

with $n = n_1 + n_2$ being the total number of parts sampled. Let $X_i = 1$ if the ith part is defective and $X_i = 0$ otherwise for $i = 1, \dots, n$. Define $Y_1 = X_1 + \cdots + X_{n_1}$ and $Y_2 = X_{n_1+1} + \cdots + X_n$. These are the total numbers of defective parts produced by each machine. The assumptions stated in the problem allow us to conclude that Y_1 and Y_2 are independent according to the note about separate functions of independent random variables on page 140. Furthermore, Theorem 3.9.2 says that Y_j has the binomial distribution with parameters n_j and p_j for $j = 1, 2$. These two marginal distributions, together with the fact that Y_1 and Y_2 are independent, give the entire joint distribution. So, for example, if g is the joint p.f. of Y_1 and Y_2, we can compute

$$g(y_1, y_2) = \binom{n_1}{y_1} p_1^{y_1}(1 - p_1)^{n_1 - y_1} \binom{n_2}{y_2} p_2^{y_2}(1 - p)^{n_2 - y_2},$$

for $y_1 = 0, \dots, n_1$ and $y_2 = 0, \dots, n_2$, while $g(y_1, y_2) = 0$ otherwise. There is no need to find a set A as in Example 3.9.2, because of the simplifying structure of the joint distribution of $X_1, \dots, X_n$. ◄

Random Variables with a Continuous Joint Distribution

Example 3.9.4

Total Service Time. Suppose that the first two customers in a queue plan to leave together. Let X_i be the time it takes to serve customer i for $i = 1, 2$. Suppose also that X_1 and X_2 are independent random variables with common distribution having p.d.f. $f(x) = 2e^{-2x}$ for $x > 0$ and 0 otherwise. Since the customers will leave together, they are interested in the total time it takes to serve both of them, namely, $Y = X_1 + X_2$. We can now find the p.d.f. of Y.

For each y, let

$$A_y = \{(x_1, x_2) : x_1 + x_2 \le y\}.$$

Then $Y \le y$ if and only if $(X_1, X_2) \in A_y$. The set A_y is pictured in Fig. 3.24. If we let $G(y)$ denote the c.d.f. of Y, then, for $y > 0$,

$$\begin{aligned} G(y) = \Pr((X_1, X_2) \in A_y) &= \int_0^y \int_0^{y-x_2} 4e^{-2x_1-2x_2} dx_1 dx_2 \\ &= \int_0^y 2e^{-2x_2}\left[1 - e^{-2(y-x_2)}\right] dx_2 = \int_0^y \left[2e^{-2x_2} - 2e^{-2y}\right] dx_2 \\ &= 1 - e^{-2y} - 2ye^{-2y}. \end{aligned}$$

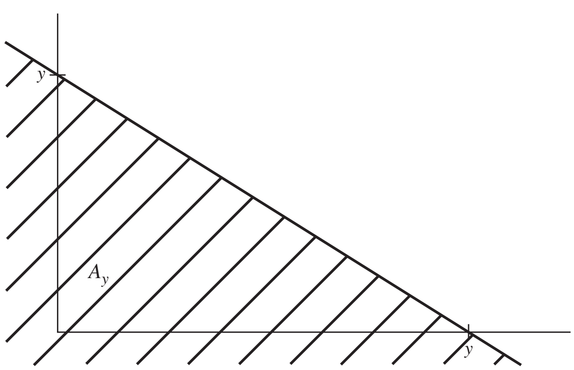


Figure 3.24 The set A_y in Example 3.9.4 and in the proof of Theorem 3.9.4.

Taking the derivative of $G(y)$ with respect to y, we get the p.d.f.

$$g(y) = \frac{d}{dy}\left[1 - e^{-2y} - ye^{-2y}\right] = 4ye^{-2y},$$

for $y > 0$ and 0 otherwise. ◀

The transformation in Example 3.9.4 is an example of a brute-force method that is always available for finding the distribution of a function of several random variables, however, it might be difficult to apply in individual cases.

Theorem 3.9.3 **Brute-Force Distribution of a Function.** Suppose that the joint p.d.f. of $\boldsymbol{X} = (X_1, \ldots, X_n)$ is $f(\boldsymbol{x})$ and that $Y = r(\boldsymbol{X})$. For each real number y, define $A_y = \{\boldsymbol{x} : r(\boldsymbol{x}) \le y\}$. Then the c.d.f. $G(y)$ of Y is

$$G(y) = \int \underset{A_y}{\cdots} \int f(\boldsymbol{x})\, d\boldsymbol{x}. \tag{3.9.1}$$

Proof From the definition of c.d.f.,

$$G(y) = \Pr(Y \le y) = \Pr[r(\boldsymbol{X}) \le y] = \Pr(\boldsymbol{X} \in A_y),$$

which equals the right side of Eq. (3.9.1) by Definition 3.7.3. ■

If the distribution of Y also is continuous, then the p.d.f. of Y can be found by differentiating the c.d.f. $G(y)$.

A popular special case of Theorem 3.9.3 is the following.

Theorem 3.9.4 **Linear Function of Two Random Variables.** Let X_1 and X_2 have joint p.d.f. $f(x_1, x_2)$, and let $Y = a_1X_1 + a_2X_2 + b$ with $a_1 \ne 0$. Then Y has a continuous distribution whose p.d.f. is

$$g(y) = \int_{-\infty}^{\infty} f\left(\frac{y - b - a_2x_2}{a_1}, x_2\right) \frac{1}{|a_1|} dx_2. \tag{3.9.2}$$

Proof First, we shall find the c.d.f. G of Y whose derivative we will see is the function g in Eq. (3.9.2). For each y, let $A_y = \{(x_1, x_2) : a_1x_1 + a_2x_2 + b \le y\}$. The set A_y has the same general form as the set in Fig. 3.24. We shall write the integral over the set A_y with x_2 in the outer integral and x_1 in the inner integral. Assume that $a_1 > 0$. The other case is similar. According to Theorem 3.9.3,

$$G(y) = \int_{A_y}\int f(x_1, x_2)dx_1dx_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{(y-b-a_2x_2)/a_1} f(x_1, x_2)dx_1dx_2. \tag{3.9.3}$$

For the inner integral, perform the change of variable $z = a_1x_1 + a_2x_2 + b$ whose inverse is $x_1 = (z - b - a_2x_2)/a_1$, so that $dx_1 = dz/a_1$. The inner integral, after this change of variable, becomes

$$\int_{-\infty}^{y} f\left(\frac{z - b - a_2x_2}{a_1}, x_2\right) \frac{1}{a_1} dz.$$

We can now substitute this expression for the inner integral into Eq. (3.9.3):

$$\begin{aligned} G(y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{y} f\left(\frac{z - b - a_2x_2}{a_1}, x_2\right) \frac{1}{a_1} dzdx_2 \\ &= \int_{-\infty}^{y}\int_{-\infty}^{\infty} f\left(\frac{z - b - a_2x_2}{a_1}, x_2\right) \frac{1}{a_1} dx_2dz. \end{aligned} \tag{3.9.4}$$

Let $g(z)$ denote the inner integral on the far right side of Eq. (3.9.4). Then we have $G(y) = \int_{-\infty}^{y} g(z)dz$, whose derivative is $g(y)$, the function in Eq. (3.9.2). ■

The special case of Theorem 3.9.4 in which X_1 and X_2 are independent, $a_1 = a_2 = 1$, and $b = 0$ is called *convolution*.

Definition 3.9.1 Convolution. Let X_1 and X_2 be independent continuous random variables and let $Y = X_1 + X_2$. The distribution of Y is called the *convolution* of the distributions of X_1 and X_2. The p.d.f. of Y is sometimes called the convolution of the p.d.f.'s of X_1 and X_2.

If we let the p.d.f. of X_i be f_i for $i = 1, 2$ in Definition 3.9.1, then Theorem 3.9.4 (with $a_1 = a_2 = 1$ and $b = 0$) says that the p.d.f. of $Y = X_1 + X_2$ is

$$g(y) = \int_{-\infty}^{\infty} f_1(y - z) f_2(z)dz. \tag{3.9.5}$$

Equivalently, by switching the names of X_1 and X_2, we obtain the alternative form for the convolution:

$$g(y) = \int_{-\infty}^{\infty} f_1(z) f_2(y - z)\, dz. \tag{3.9.6}$$

The p.d.f. found in Example 3.9.4 is the special case of (3.9.5) with $f_1(x) = f_2(x) = 2e^{-2x}$ for $x > 0$ and 0 otherwise.

Example 3.9.5 An Investment Portfolio. Suppose that an investor wants to purchase both stocks and bonds. Let X_1 be the value of the stocks at the end of one year, and let X_2 be the value of the bonds at the end of one year. Suppose that X_1 and X_2 are independent. Let X_1 have the uniform distribution on the interval [1000, 4000], and let X_2 have the uniform distribution on the interval [800, 1200]. The sum $Y = X_1 + X_2$ is the value at the end of the year of the portfolio consisting of both the stocks and the bonds. We shall find the p.d.f. of Y. The function $f_1(z) f_2(y - z)$ in Eq. (3.9.6) is

$$f_1(z) f_2(y - z) = \begin{cases} 8.333 \times 10^{-7} & \text{for } 1000 \le z \le 4000 \\ & \quad \text{and } 800 \le y - z \le 1200, \\ 0 & \text{otherwise.} \end{cases} \tag{3.9.7}$$

We need to integrate the function in Eq. (3.9.7) over z for each value of y to get the marginal p.d.f. of Y. It is helpful to look at a graph of the set of (y, z) pairs for which the function in Eq. (3.9.7) is positive. Figure 3.25 shows the region shaded. For $1800 < y \le 2200$, we must integrate z from 1000 to $y - 800$. For $2200 < y \le 4800$, we must integrate z from $y - 1200$ to $y - 800$. For $4800 < y < 5200$, we must integrate z from $y - 1200$ to 4000. Since the function in Eq. (3.9.7) is constant when it is positive, the integral equals the constant times the length of the interval of z values. So, the p.d.f. of Y is

$$g(y) = \begin{cases} 8.333 \times 10^{-7}(y - 1800) & \text{for } 1800 < y \le 2200, \\ 3.333 \times 10^{-4} & \text{for } 2200 < y \le 4800, \\ 8.333 \times 10^{-7}(5200 - y) & \text{for } 4800 < y < 5200, \\ 0 & \text{otherwise.} \end{cases}$$ ◄

As another example of the brute-force method, we consider the largest and smallest observations in a random sample. These functions give an idea of how spread out the sample is. For example, meteorologists often report record high and low

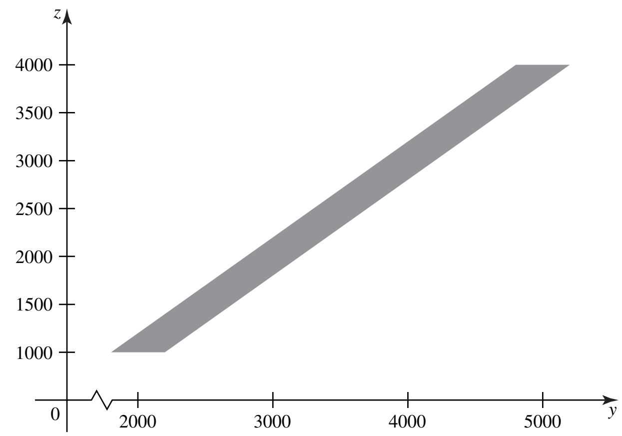


Figure 3.25 The region where the function in Eq. (3.9.7) is positive.

temperatures for specific days as well as record high and low rainfalls for months and years.

Example 3.9.6

Maximum and Minimum of a Random Sample. Suppose that $X_1, \dots, X_n$ form a random sample of size n from a distribution for which the p.d.f. is f and the c.d.f. is F. The largest value Y_n and the smallest value Y_1 in the random sample are defined as follows:

$$\begin{aligned} Y_n &= \max\{X_1, \dots, X_n\}, \\ Y_1 &= \min\{X_1, \dots, X_n\}. \end{aligned} \tag{3.9.8}$$

Consider Y_n first. Let G_n stand for its c.d.f., and let g_n be its p.d.f. For every given value of y $(-\infty < y < \infty)$,

$$\begin{aligned} G_n(y) &= \Pr(Y_n \le y) = \Pr(X_1 \le y, X_2 \le y, \dots, X_n \le y) \\ &= \Pr(X_1 \le y)\Pr(X_2 \le y) \cdots \Pr(X_n \le y) \\ &= F(y)F(y) \cdots F(y) = [F(y)]^n, \end{aligned}$$

where the third equality follows from the fact that the X_i are independent and the fourth follows from the fact that all of the X_i have the same c.d.f. F. Thus, $G_n(y) = [F(y)]^n$.

Now, g_n can be determined by differentiating the c.d.f. G_n. The result is

$$g_n(y) = n[F(y)]^{n-1} f(y) \quad \text{for } -\infty < y < \infty.$$

Next, consider Y_1 with c.d.f. G_1 and p.d.f. g_1. For every given value of y $(-\infty < y < \infty)$,

$$\begin{aligned} G_1(y) &= \Pr(Y_1 \le y) = 1 - \Pr(Y_1 > y) \\ &= 1 - \Pr(X_1 > y, X_2 > y, \dots, X_n > y) \\ &= 1 - \Pr(X_1 > y)\Pr(X_2 > y) \cdots \Pr(X_n > y) \\ &= 1 - [1 - F(y)][1 - F(y)] \cdots [1 - F(y)] \\ &= 1 - [1 - F(y)]^n. \end{aligned}$$

Thus, $G_1(y) = 1 - [1 - F(y)]^n$.

Then g_1 can be determined by differentiating the c.d.f. G_1. The result is

$$g_1(y) = n[1 - F(y)]^{n-1} f(y) \quad \text{for } -\infty < y < \infty.$$

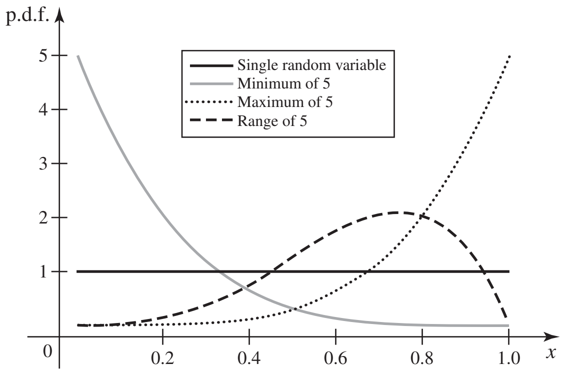


Figure 3.26 The p.d.f. of the uniform distribution on the interval [0, 1] together with the p.d.f.'s of the minimum and maximum of samples of size $n = 5$. The p.d.f. of the range of a sample of size $n = 5$ (see Example 3.9.7) is also included.

Figure 3.26 shows the p.d.f. of the uniform distribution on the interval [0, 1] together with the p.d.f.'s of Y_1 and Y_n for the case $n = 5$. It also shows the p.d.f. of $Y_5 - Y_1$, which will be derived in Example 3.9.7. Notice that the p.d.f. of Y_1 is highest near 0 and lowest near 1, while the opposite is true of the p.d.f. of Y_n, as one would expect.

Finally, we shall determine the joint distribution of Y_1 and Y_n. For every pair of values (y_1, y_n) such that $-\infty < y_1 < y_n < \infty$, the event $\{Y_1 \le y_1\} \cap \{Y_n \le y_n\}$ is the same as $\{Y_n \le y_n\} \cap \{Y_1 > y_1\}^c$. If G denotes the bivariate joint c.d.f. of Y_1 and Y_n, then

$$\begin{aligned}
G(y_1, y_n) &= \Pr(Y_1 \le y_1 \text{ and } Y_n \le y_n)\\
&= \Pr(Y_n \le y_n) - \Pr(Y_n \le y_n \text{ and } Y_1 > y_1)\\
&= \Pr(Y_n \le y_n)\\
&\quad - \Pr(y_1 < X_1 \le y_n, y_1 < X_2 \le y_n, \ldots, y_1 < X_n \le y_n)\\
&= G_n(y_n) - \prod_{i=1}^{n} \Pr(y_1 < X_i \le y_n)\\
&= [F(y_n)]^n - [F(y_n) - F(y_1)]^n.
\end{aligned}$$

The bivariate joint p.d.f. g of Y_1 and Y_n can be found from the relation

$$g(y_1, y_n) = \frac{\partial^2 G(y_1, y_n)}{\partial y_1 \partial y_n}.$$

Thus, for $-\infty < y_1 < y_n < \infty$,

$$g(y_1, y_n) = n(n-1)[F(y_n) - F(y_1)]^{n-2} f(y_1) f(y_n). \tag{3.9.9}$$

Also, for all other values of y_1 and y_n, $g(y_1, y_n) = 0$. ◀

A popular way to describe how spread out is a random sample is to use the distance from the minimum to the maximum, which is called the *range* of the random sample. We can combine the result from the end of Example 3.9.6 with Theorem 3.9.4 to find the p.d.f. of the range.

Example 3.9.7 **The Distribution of the Range of a Random Sample.** Consider the same situation as in Example 3.9.6. The random variable $W = Y_n - Y_1$ is called the *range* of the sample. The joint p.d.f. $g(y_1, y_n)$ of Y_1 and Y_n was presented in Eq. (3.9.9). We can now apply Theorem 3.9.4 with $a_1 = -1$, $a_2 = 1$, and $b = 0$ to get the p.d.f. h of W:

$$h(w) = \int_{-\infty}^{\infty} g(y_n - w, y_n)dy_n = \int_{-\infty}^{\infty} g(z, z + w)dz, \tag{3.9.10}$$

where, for the last equality, we have made the change of variable $z = y_n - w$. ◀

Here is a special case in which the integral of Eq. 3.9.10 can be computed in closed form.

Example 3.9.8

The Range of a Random Sample from a Uniform Distribution. Suppose that the n random variables $X_1, \dots, X_n$ form a random sample from the uniform distribution on the interval [0, 1]. We shall determine the p.d.f. of the range of the sample.

In this example,

$$f(x) = \begin{cases} 1 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

Also, $F(x) = x$ for $0 < x < 1$. We can write $g(y_1, y_n)$ from Eq. (3.9.9) in this case as

$$g(y_1, y_n) = \begin{cases} n(n-1)(y_n - y_1)^{n-2} & \text{for } 0 < y_1 < y_n < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, in Eq. (3.9.10), $g(z, z + w) = 0$ unless $0 < w < 1$ and $0 < z < 1 - w$. For values of w and z satisfying these conditions, $g(z, w + z) = n(n - 1)w^{n-2}$. The p.d.f. in Eq. (3.9.10) is then, for $0 < w < 1$,

$$h(w) = \int_0^{1-w} n(n-1)w^{n-2}\, dz = n(n-1)w^{n-2}(1-w).$$

Otherwise, $h(w) = 0$. This p.d.f. is shown in Fig. 3.26 for the case $n = 5$. ◀

Direct Transformation of a Multivariate p.d.f.

Next, we state without proof a generalization of Theorem 3.8.4 to the case of several random variables. The proof of Theorem 3.9.5 is based on the theory of differentiable one-to-one transformations in advanced calculus.

Theorem 3.9.5

Multivariate Transformation. Let $X_1, \dots, X_n$ have a continuous joint distribution for which the joint p.d.f. is f. Assume that there is a subset S of R^n such that $\Pr[(X_1, \dots, X_n) \in S] = 1$. Define n new random variables $Y_1, \dots, Y_n$ as follows:

$$\begin{aligned} Y_1 &= r_1(X_1, \dots, X_n), \\ Y_2 &= r_2(X_1, \dots, X_n), \\ &\vdots \\ Y_n &= r_n(X_1, \dots, X_n), \end{aligned} \tag{3.9.11}$$

where we assume that the n functions $r_1, \dots, r_n$ define a one-to-one differentiable transformation of S onto a subset T of R^n. Let the inverse of this transformation be given as follows:

$$\begin{aligned} x_1 &= s_1(y_1, \dots, y_n), \\ x_2 &= s_2(y_1, \dots, y_n), \\ &\vdots \\ x_n &= s_n(y_1, \dots, y_n). \end{aligned} \tag{3.9.12}$$

Then the joint p.d.f. g of $Y_1, \dots, Y_n$ is

$$g(y_1, \dots, y_n) = \begin{cases} f(s_1, \dots, s_n)|J| & \text{for } (y_1, \dots, y_n) \in T, \\ 0 & \text{otherwise,} \end{cases} \tag{3.9.13}$$

where J is the determinant

$$J = \det \begin{bmatrix} \frac{\partial s_1}{\partial y_1} & \cdots & \frac{\partial s_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_n}{\partial y_1} & \cdots & \frac{\partial s_n}{\partial y_n} \end{bmatrix}$$

and $|J|$ denotes the absolute value of the determinant J. ∎

Thus, the joint p.d.f. $g(y_1, \dots, y_n)$ is obtained by starting with the joint p.d.f. $f(x_1, \dots, x_n)$, replacing each value x_i by its expression $s_i(y_1, \dots, y_n)$ in terms of $y_1, \dots, y_n$, and then multiplying the result by $|J|$. This determinant J is called the *Jacobian* of the transformation specified by the equations in (3.9.12).

Note: The Jacobian Is a Generalization of the Derivative of the Inverse. Eqs. (3.8.3) and (3.9.13) are very similar. The former gives the p.d.f. of a single function of a single random variable. Indeed, if $n = 1$ in (3.9.13), $J = ds_1(y_1)/dy_1$ and Eq. (3.9.13) becomes the same as (3.8.3). The Jacobian merely generalizes the derivative of the inverse of a single function of one variable to n functions of n variables.

Example 3.9.9 **The Joint p.d.f. of the Quotient and the Product of Two Random Variables.** Suppose that two random variables X_1 and X_2 have a continuous joint distribution for which the joint p.d.f. is as follows:

$$f(x_1, x_2) = \begin{cases} 4x_1x_2 & \text{for } 0 < x_1 < 1 \text{ and } 0 < x_2 < 1, \\ 0 & \text{otherwise.} \end{cases}$$

We shall determine the joint p.d.f. of two new random variables Y_1 and Y_2, which are defined by the relations

$$Y_1 = \frac{X_1}{X_2} \text{ and } Y_2 = X_1X_2.$$

In the notation of Theorem 3.9.5, we would say that $Y_1 = r_1(X_1, X_2)$ and $Y_2 = r_2(X_1, X_2)$, where

$$r_1(x_1, x_2) = \frac{x_1}{x_2} \text{ and } r_2(x_1, x_2) = x_1x_2. \tag{3.9.14}$$

The inverse of the transformation in Eq. (3.9.14) is found by solving the equations $y_1 = r_1(x_1, x_2)$ and $y_2 = r_2(x_1, x_2)$ for x_1 and x_2 in terms of y_1 and y_2. The result is

$$\begin{aligned} x_1 &= s_1(y_1, y_2) = (y_1y_2)^{1/2}, \\ x_2 &= s_2(y_1, y_2) = \left(\frac{y_2}{y_1}\right)^{1/2}. \end{aligned} \tag{3.9.15}$$

Let S denote the set of points (x_1, x_2) such that $0 < x_1 < 1$ and $0 < x_2 < 1$, so that $\Pr[(X_1, X_2) \in S] = 1$. Let T be the set of (y_1, y_2) pairs such that $(y_1, y_2) \in T$ if and only if $(s_1(y_1, y_2), s_2(y_1, y_2)) \in S$. Then $\Pr[(Y_1, Y_2) \in T] = 1$. The transformation defined by the equations in (3.9.14) or, equivalently, by the equations in (3.9.15) specifies a one-to-one relation between the points in S and the points in T.

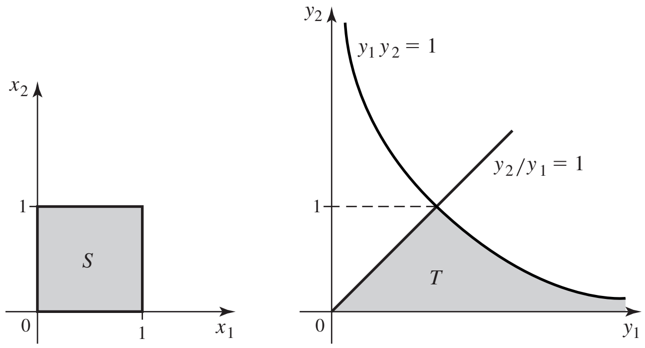


Figure 3.27 The sets S and T in Example 3.9.9.

We shall now show how to find the set T. We know that $(x_1, x_2) \in S$ if and only if the following inequalities hold:

$$x_1 > 0, \quad x_1 < 1, \quad x_2 > 0, \text{ and } x_2 < 1. \tag{3.9.16}$$

We can substitute the formulas for x_1 and x_2 in terms of y_1 and y_2 from Eq. (3.9.15) into the inequalities in (3.9.16) to obtain

$$(y_1 y_2)^{1/2} > 0, \quad (y_1 y_2)^{1/2} < 1, \quad \left(\frac{y_2}{y_1}\right)^{1/2} > 0,$$
$$\text{and } \left(\frac{y_2}{y_1}\right)^{1/2} < 1. \tag{3.9.17}$$

The first inequality transforms to ($y_1 > 0$ and $y_2 > 0$) or ($y_1 < 0$ and $y_2 < 0$). However, since $y_1 = x_1/x_2$, we cannot have $y_1 < 0$, so we get only $y_1 > 0$ and $y_2 > 0$. The third inequality in (3.9.17) transforms to the same thing. The second inequality in (3.9.17) becomes $y_2 < 1/y_1$. The fourth inequality becomes $y_2 < y_1$. The region T where (y_1, y_2) satisfy these new inequalities is shown in the right panel of Fig. 3.27 with the set S in the left panel.

For the functions in (3.9.15),

$$\frac{\partial s_1}{\partial y_1} = \frac{1}{2}\left(\frac{y_2}{y_1}\right)^{1/2}, \quad \frac{\partial s_1}{\partial y_2} = \frac{1}{2}\left(\frac{y_1}{y_2}\right)^{1/2},$$
$$\frac{\partial s_2}{\partial y_1} = -\frac{1}{2}\left(\frac{y_2}{y_1^3}\right)^{1/2}, \quad \frac{\partial s_2}{\partial y_2} = \frac{1}{2}\left(\frac{1}{y_1 y_2}\right)^{1/2}.$$

Hence,

$$J = \det\begin{bmatrix} \frac{1}{2}\left(\frac{y_2}{y_1}\right)^{1/2} & \frac{1}{2}\left(\frac{y_1}{y_2}\right)^{1/2} \\ -\frac{1}{2}\left(\frac{y_2}{y_1^3}\right)^{1/2} & \frac{1}{2}\left(\frac{1}{y_1 y_2}\right)^{1/2} \end{bmatrix} = \frac{1}{2y_1}.$$

Since $y_1 > 0$ throughout the set T, $|J| = 1/(2y_1)$.

The joint p.d.f. $g(y_1, y_2)$ can now be obtained directly from Eq. (3.9.13) in the following way: In the expression for $f(x_1, x_2)$, replace x_1 with $(y_1 y_2)^{1/2}$, replace x_2

with $(y_2/y_1)^{1/2}$, and multiply the result by $|J| = 1/(2y_1)$. Therefore,

$$g(y_1, y_2) = \begin{cases} 2\left(\frac{y_2}{y_1}\right) & \text{for } (y_1, y_2) \in T, \\ 0 & \text{otherwise.} \end{cases}$$ ◀

Example 3.9.10

Service Time in a Queue. Let X be the time that the server in a single-server queue will spend on a particular customer, and let Y be the rate at which the server can operate. A popular model for the conditional distribution of X given $Y = y$ is to say that the conditional p.d.f. of X given $Y = y$ is

$$g_1(x|y) = \begin{cases} ye^{-xy} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let Y have the p.d.f. $f_2(y)$. The joint p.d.f. of (X, Y) is then $g_1(x|y)f_2(y)$. Because $1/Y$ can be interpreted as the average service time, $Z = XY$ measures how quickly, compared to average, that the customer is served. For example, $Z = 1$ corresponds to an average service time, while $Z > 1$ means that this customer took longer than average, and $Z < 1$ means that this customer was served more quickly than the average customer. If we want the distribution of Z, we could compute the joint p.d.f. of (Z, Y) directly using the methods just illustrated. We could then integrate the joint p.d.f. over y to obtain the marginal p.d.f. of Z. However, it is simpler to transform the conditional distribution of X given $Y = y$ into the conditional distribution of Z given $Y = y$, since conditioning on $Y = y$ allows us to treat Y as the constant y. Because $X = Z/Y$, the inverse transformation is $x = s(z)$, where $s(z) = z/y$. The derivative of this is $1/y$, and the conditional p.d.f. of Z given $Y = y$ is

$$h_1(z|y) = \frac{1}{y} g_1\left(\frac{z}{y}\,\middle|\, y\right).$$

Because Y is a rate, $Y \ge 0$ and $X = Z/Y > 0$ if and only if $Z > 0$. So,

$$h_1(z|y) = \begin{cases} e^{-z} & \text{for } z > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.9.18}$$

Notice that h_1 does not depend on y, so Z is independent of Y and h_1 is the marginal p.d.f. of Z. The reader can verify all of this in Exercise 17. ◀

Note: Removing Dependence. The formula $Z = XY$ in Example 3.9.10 makes it look as if Z should depend on Y. In reality, however, multiplying X by Y removes the dependence that X already has on Y and makes the result independent of Y. This type of transformation that removes the dependence of one random variable on another is a very powerful technique for finding marginal distributions of transformations of random variables.

In Example 3.9.10, we mentioned that there was another, more straightforward but more tedious, way to compute the distribution of Z. That method, which is useful in many settings, is to transform (X, Y) into (Z, W) for some uninteresting random variable W and then integrate w out of the joint p.d.f. All that matters in the choice of W is that the transformation be one-to-one with differentiable inverse and that the calculations are feasible. Here is a specific example.

Example 3.9.11

One Function of Two Variables. In Example 3.9.9, suppose that we were interested only in the quotient $Y_1 = X_1/X_2$ rather than both the quotient and the product $Y_2 = X_1X_2$. Since we already have the joint p.d.f. of (Y_1, Y_2), we will merely integrate y_2 out rather than start from scratch. For each value of $y_1 > 0$, we need to look at the set T in Fig. 3.27 and find the interval of y_2 values to integrate over. For $0 < y_1 < 1$,

we integrate over $0 < y_2 < y_1$. For $y_1 > 1$, we integrate over $0 < y_2 < 1/y_1$. (For $y_1 = 1$ both intervals are the same.) So, the marginal p.d.f. of Y_1 is

$$g_1(y_1) = \begin{cases} \int_0^{y_1} 2\left(\frac{y_2}{y_1}\right) dy_2 & \text{for } 0 < y_1 < 1, \\ \int_0^{1/y_1} 2\left(\frac{y_2}{y_1}\right) dy_2 & \text{for } y_1 > 1, \end{cases}$$
$$= \begin{cases} y_1 & \text{for } 0 < y_1 < 1, \\ \frac{1}{y_1^3} & \text{for } y_1 > 1. \end{cases}$$

There are other transformations that would have made the calculation of g_1 simpler if that had been all we wanted. See Exercise 21 for an example. ◀

Theorem 3.9.6 Linear Transformations. Let $\boldsymbol{X} = (X_1, \dots, X_n)$ have a continuous joint distribution for which the joint p.d.f. is f. Define $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ by

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}, \tag{3.9.19}$$

where $\boldsymbol{A}$ is a nonsingular $n \times n$ matrix. Then $\boldsymbol{Y}$ has a continuous joint distribution with p.d.f.

$$g(\boldsymbol{y}) = \frac{1}{|\det \boldsymbol{A}|} f(\boldsymbol{A}^{-1}\boldsymbol{y}) \quad \text{for } \boldsymbol{y} \in R^n, \tag{3.9.20}$$

where $\boldsymbol{A}^{-1}$ is the inverse of $\boldsymbol{A}$.

Proof Each Y_i is a linear combination of $X_1, \dots, X_n$. Because $\boldsymbol{A}$ is nonsingular, the transformation in Eq. (3.9.19) is a one-to-one transformation of the entire space R^n onto itself. At every point $\boldsymbol{y} \in R^n$, the inverse transformation can be represented by the equation

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}. \tag{3.9.21}$$

The Jacobian J of the transformation that is defined by Eq. (3.9.21) is simply $J = \det \boldsymbol{A}^{-1}$. Also, it is known from the theory of determinants that

$$\det \boldsymbol{A}^{-1} = \frac{1}{\det \boldsymbol{A}}.$$

Therefore, at every point $\boldsymbol{y} \in R^n$, the joint p.d.f. $g(\boldsymbol{y})$ can be evaluated in the following way, according to Theorem 3.9.5: First, for $i = 1, \dots, n$, the component x_i in $f(x_1, \dots, x_n)$ is replaced with the ith component of the vector $\boldsymbol{A}^{-1}\boldsymbol{y}$. Then, the result is divided by $|\det \boldsymbol{A}|$. This produces Eq. (3.9.20). ■

Summary

We extended the construction of the distribution of a function of a random variable to the case of several functions of several random variables. If one only wants the distribution of one function r_1 of n random variables, the usual way to find this is to first find $n - 1$ additional functions $r_2, \dots, r_n$ so that the n functions together compose a one-to-one transformation. Then find the joint p.d.f. of the n functions and finally find the marginal p.d.f. of the first function by integrating out the extra $n - 1$ variables. The method is illustrated for the cases of the sum and the range of several random variables.

Exercises

1. Suppose that X_1 and X_2 are i.i.d. random variables and that each of them has the uniform distribution on the interval [0, 1]. Find the p.d.f. of $Y = X_1 + X_2$.

2. For the conditions of Exercise 1, find the p.d.f. of the average $(X_1 + X_2)/2$.

3. Suppose that three random variables X_1, X_2, and X_3 have a continuous joint distribution for which the joint p.d.f. is as follows:

$$f(x_1, x_2, x_3) = \begin{cases} 8x_1x_2x_3 & \text{for } 0 < x_i < 1 \ \ (i = 1, 2, 3), \\ 0 & \text{otherwise.} \end{cases}$$

Suppose also that $Y_1 = X_1$, $Y_2 = X_1X_2$, and $Y_3 = X_1X_2X_3$. Find the joint p.d.f. of Y_1, Y_2, and Y_3.

4. Suppose that X_1 and X_2 have a continuous joint distribution for which the joint p.d.f. is as follows:

$$f(x_1, x_2) = \begin{cases} x_1 + x_2 & \text{for } 0 < x_1 < 1 \text{ and } 0 < x_2 < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the p.d.f. of $Y = X_1X_2$.

5. Suppose that the joint p.d.f. of X_1 and X_2 is as given in Exercise 4. Find the p.d.f. of $Z = X_1/X_2$.

6. Let X and Y be random variables for which the joint p.d.f. is as follows:

$$f(x, y) = \begin{cases} 2(x + y) & \text{for } 0 \le x \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the p.d.f. of $Z = X + Y$.

7. Suppose that X_1 and X_2 are i.i.d. random variables and that the p.d.f. of each of them is as follows:

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Find the p.d.f. of $Y = X_1 - X_2$.

8. Suppose that $X_1, \ldots, X_n$ form a random sample of size n from the uniform distribution on the interval [0, 1] and that $Y_n = \max\{X_1, \ldots, X_n\}$. Find the smallest value of n such that

$$\Pr\{Y_n \ge 0.99\} \ge 0.95.$$

9. Suppose that the n variables $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval [0, 1] and that the random variables Y_1 and Y_n are defined as in Eq. (3.9.8). Determine the value of $\Pr(Y_1 \le 0.1$ and $Y_n \le 0.8)$.

10. For the conditions of Exercise 9, determine the value of $\Pr(Y_1 \le 0.1$ and $Y_n \ge 0.8)$.

11. For the conditions of Exercise 9, determine the probability that the interval from Y_1 to Y_n will not contain the point 1/3.

12. Let W denote the range of a random sample of n observations from the uniform distribution on the interval [0, 1]. Determine the value of $\Pr(W > 0.9)$.

13. Determine the p.d.f. of the range of a random sample of n observations from the uniform distribution on the interval [−3, 5].

14. Suppose that $X_1, \ldots, X_n$ form a random sample of n observations from the uniform distribution on the interval [0, 1], and let Y denote the second largest of the observations. Determine the p.d.f. of Y. *Hint:* First determine the c.d.f. G of Y by noting that

$$\begin{aligned} G(y) &= \Pr(Y \le y) \\ &= \Pr(\text{At least } n - 1 \text{ observations} \le y). \end{aligned}$$

15. Show that if $X_1, X_2, \ldots, X_n$ are independent random variables and if $Y_1 = r_1(X_1), Y_2 = r_2(X_2), \ldots, Y_n = r_n(X_n)$, then $Y_1, Y_2, \ldots, Y_n$ are also independent random variables.

16. Suppose that $X_1, X_2, \ldots, X_5$ are five random variables for which the joint p.d.f. can be factored in the following form for all points $(x_1, x_2, \ldots, x_5) \in R^5$:

$$f(x_1, x_2, \ldots, x_5) = g(x_1, x_2)h(x_3, x_4, x_5),$$

where g and h are certain nonnegative functions. Show that if $Y_1 = r_1(X_1, X_2)$ and $Y_2 = r_2(X_3, X_4, X_5)$, then the random variables Y_1 and Y_2 are independent.

17. In Example 3.9.10, use the Jacobian method (3.9.13) to verify that Y and Z are independent and that Eq. (3.9.18) is the marginal p.d.f. of Z.

18. Let the conditional p.d.f. of X given Y be $g_1(x|y) = 3x^2/y^3$ for $0 < x < y$ and 0 otherwise. Let the marginal p.d.f. of Y be $f_2(y)$, where $f_2(y) = 0$ for $y \le 0$ but is otherwise unspecified. Let $Z = X/Y$. Prove that Z and Y are independent and find the marginal p.d.f. of Z.

19. Let X_1 and X_2 be as in Exercise 7. Find the p.d.f. of $Y = X_1 + X_2$.

20. If $a_2 = 0$ in Theorem 3.9.4, show that Eq. (3.9.2) becomes the same as Eq. (3.8.1) with $a = a_1$ and $f = f_1$.

21. In Examples 3.9.9 and 3.9.11, find the marginal p.d.f. of $Z_1 = X_1/X_2$ by first transforming to Z_1 and $Z_2 = X_1$ and then integrating z_2 out of the joint p.d.f.

★ 3.10 Markov Chains

A popular model for systems that change over time in a random manner is the Markov chain model. A Markov chain is a sequence of random variables, one for each time. At each time, the corresponding random variable gives the state of the system. Also, the conditional distribution of each future state given the past states and the present state depends only on the present state.

Stochastic Processes

Example 3.10.1 Occupied Telephone Lines. Suppose that a certain business office has five telephone lines and that any number of these lines may be in use at any given time. During a certain period of time, the telephone lines are observed at regular intervals of 2 minutes and the number of lines that are being used at each time is noted. Let X_1 denote the number of lines that are being used when the lines are first observed at the beginning of the period; let X_2 denote the number of lines that are being used when they are observed the second time, 2 minutes later; and in general, for $n = 1, 2, \ldots,$ let X_n denote the number of lines that are being used when they are observed for the nth time. ◀

Definition 3.10.1 Stochastic Process. A sequence of random variables $X_1, X_2, \ldots$ is called a *stochastic process* or *random process* with *discrete time parameter*. The first random variable X_1 is called the *initial state* of the process; and for $n = 2, 3, \ldots,$ the random variable X_n is called the *state of the process at time n*.

In Example 3.10.1, the state of the process at any time is the number of lines being used at that time. Therefore, each state must be an integer between 0 and 5. Each of the random variables in a stochastic process has a marginal distribution, and the entire process has a joint distribution. For convenience, in this text, we will discuss only joint distributions for finitely many of $X_1, X_2, \ldots$ at a time. The meaning of the phrase "discrete time parameter" is that the process, such as the numbers of occupied phone lines, is observed only at discrete or separated points in time, rather than continuously in time. In Sec. 5.4, we will introduce a different stochastic process (called the Poisson process) with a continuous time parameter.

In a stochastic process with a discrete time parameter, the state of the process varies in a random manner from time to time. To describe a complete probability model for a particular process, it is necessary to specify the distribution for the initial state X_1 and also to specify for each $n = 1, 2, \ldots$ the conditional distribution of the subsequent state X_{n+1} given $X_1, \ldots, X_n$. These conditional distributions are equivalent to the collection of conditional c.d.f.'s of the following form:

$$\Pr(X_{n+1} \le b | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Markov Chains

A Markov chain is a special type of stochastic process, defined in terms of the conditional distributions of future states given the present and past states.

Definition 3.10.2 Markov Chain. A stochastic process with discrete time parameter is a *Markov chain* if, for each time n, the conditional distributions of all X_{n+j} for $j \ge 1$ given $X_1, \ldots, X_n$ depend only on X_n and not on the earlier states $X_1, \ldots, X_{n-1}$. In symbols, for

$n = 1, 2, \ldots$ and for each b and each possible sequence of states $x_1, x_2, \ldots, x_n$,

$$\Pr(X_{n+1} \le b | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_{n+1} \le b | X_n = x_n).$$

A Markov chain is called *finite* if there are only finitely many possible states.

In the remainder of this section, we shall consider only finite Markov chains. This assumption could be relaxed at the cost of more complicated theory and calculation. For convenience, we shall reserve the symbol k to stand for the number of possible states of a general finite Markov chain for the remainder of the section. It will also be convenient, when discussing a general finite Markov chain, to name the k states using the integers $1, \ldots, k$. That is, for each n and j, $X_n = j$ will mean that the chain is in state j at time n. In specific examples, it may prove more convenient to label the states in a more informative fashion. For example, if the states are the numbers of phone lines in use at given times (as in the example that introduced this section), we would label the states $0, \ldots, 5$ even though $k = 6$.

The following result follows from the multiplication rule for conditional probabilities, Theorem 2.1.2.

Theorem 3.10.1 For a finite Markov chain, the joint p.f. for the first n states equals

$$\begin{aligned}\Pr&(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)\\ &= \Pr(X_1 = x_1)\Pr(X_2 = x_2|X_1 = x_1)\Pr(X_3 = x_3|X_2 = x_2)\cdots\\ &\quad \Pr(X_n = x_n|X_{n-1} = x_{n-1}).\end{aligned} \tag{3.10.1}$$

Also, for each n and each $m > 0$,

$$\begin{aligned}\Pr&(X_{n+1} = x_{n+1}, X_{n+2} = x_{n+2}, \ldots, X_{n+m} = x_{n+m}|X_n = x_n)\\ &= \Pr(X_{n+1} = x_{n+1}|X_n = x_n)\Pr(X_{n+2} = x_{n+2}|X_{n+1} = x_{n+1})\\ &\quad \cdots \Pr(X_{n+m} = x_{n+m}|X_{n+m-1} = x_{n+m-1}).\end{aligned} \tag{3.10.2}$$

■

Eq. (3.10.1) is a discrete version of a generalization of conditioning in sequence that was illustrated in Example 3.7.18 with continuous random variables. Eq. (3.10.2) is a conditional version of (3.10.1) shifted forward in time.

Example 3.10.2 Shopping for Toothpaste. In Exercise 4 in Sec. 2.1, we considered a shopper who chooses between two brands of toothpaste on several occasions. Let $X_i = 1$ if the shopper chooses brand A on the ith purchase, and let $X_i = 2$ if the shopper chooses brand B on the ith purchase. Then the sequence of states $X_1, X_2, \ldots$ is a stochastic process with two possible states at each time. The probabilities of purchase were specified by saying that the shopper will choose the same brand as on the previous purchase with probability 1/3 and will switch with probability 2/3. Since this happens regardless of purchases that are older than the previous one, we see that this stochastic process is a Markov chain with

$$\Pr(X_{n+1} = 1|X_n = 1) = \frac{1}{3},\ \Pr(X_{n+1} = 2|X_n = 1) = \frac{2}{3},$$
$$\Pr(X_{n+1} = 1|X_n = 2) = \frac{2}{3},\ \Pr(X_{n+1} = 2|X_n = 2) = \frac{1}{3}.$$ ◄

Exammple 3.10.2 has an additional feature that puts it in a special class of Markov chains. The probability of moving from one state at time n to another state at time $n + 1$ does not depend on n.

Definition 3.10.3 Transition Distributions/Stationary Transition Distributions. Consider a finite Markov chain with k possible states. The conditional distributions of the state at time $n+1$ given the state at time n, that is, $\Pr(X_{n+1} = j | X_n = i)$ for $i, j = 1, \ldots, k$ and $n = 1, 2, \ldots$, are called the *transition distributions* of the Markov chain. If the transition distribution is the same for every time n ($n = 1, 2, \ldots$), then the Markov chain has *stationary transition distributions*.

When a Markov chain with k possible states has stationary transition distributions, there exist probabilities p_{ij} for $i, j = 1, \ldots, k$ such that, for all n,

$$\Pr(X_{n+1} = j | X_n = i) = p_{ij} \quad \text{for } n = 1, 2, \ldots. \tag{3.10.3}$$

The Markov chain in Example 3.10.2 has stationary transition distributions. For example, $p_{11} = 1/3$.

In the language of multivariate distributions, when a Markov chain has stationary transition distributions, specified by (3.10.3), we can write the conditional p.f. of X_{n+1} given X_n as

$$g(j|i) = p_{ij}, \tag{3.10.4}$$

for all n, i, j.

Example 3.10.3 Occupied Telephone Lines. To illustrate the application of these concepts, we shall consider again the example involving the office with five telephone lines. In order for this stochastic process to be a Markov chain, the specified distribution for the number of lines that may be in use at each time must depend only on the number of lines that were in use when the process was observed most recently 2 minutes earlier and must not depend on any other observed values previously obtained. For example, if three lines were in use at time n, then the distribution for time $n+1$ must be the same regardless of whether 0, 1, 2, 3, 4, or 5 lines were in use at time $n-1$. In reality, however, the observation at time $n-1$ might provide some information in regard to the length of time for which each of the three lines in use at time n had been occupied, and this information might be helpful in determining the distribution for time $n+1$. Nevertheless, we shall suppose now that this process is a Markov chain. If this Markov chain is to have stationary transition distributions, it must be true that the rates at which incoming and outgoing telephone calls are made and the average duration of these telephone calls do not change during the entire period covered by the process. This requirement means that the overall period cannot include busy times when more calls are expected or quiet times when fewer calls are expected. For example, if only one line is in use at a particular observation time, regardless of when this time occurs during the entire period covered by the process, then there must be a specific probability p_{1j} that exactly j lines will be in use 2 minutes later. ◀

The Transition Matrix

Example 3.10.4 Shopping for Toothpaste. The notation for stationary transition distributions, p_{ij}, suggests that they could be arranged in a matrix. The transition probabilities for Example 3.10.2 can be arranged into the following matrix:

$$\boldsymbol{P} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}.$$ ◀

Every finite Markov chain with stationary transition distributions has a matrix like the one constructed in Example 3.10.4.

Definition 3.10.4 Transition Matrix. Consider a finite Markov chain with stationary transition distributions given by $p_{ij} = \Pr(X_{n+1} = j | X_n = i)$ for all n, i, j. The *transition matrix* of the Markov chain is defined to be the $k \times k$ matrix $\boldsymbol{P}$ with elements p_{ij}. That is,

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & \cdots & p_{1k} \\ p_{21} & \cdots & p_{2k} \\ \vdots & \ddots & \vdots \\ p_{k1} & \cdots & p_{kk} \end{bmatrix}. \tag{3.10.5}$$

A transition matrix has several properties that are apparent from its defintion. For example, each element is nonnegative because all elements are probabilities. Since each row of a transition matrix is a conditional p.f. for the next state given some value of the current state, we have $\sum_{j=1}^{k} p_{ij} = 1$ for $i = 1, \ldots, k$. Indeed, row i of the transition matrix specifies the conditional p.f. $g(\cdot|i)$ defined in (3.10.4).

Definition 3.10.5 Stochastic Matrix. A square matrix for which all elements are nonnegative and the sum of the elements in each row is 1 is called a *stochastic matrix*.

It is clear that the transition matrix $\boldsymbol{P}$ for every finite Markov chain with stationary transition probabilities must be a stochastic matrix. Conversely, every $k \times k$ stochastic matrix can serve as the transition matrix of a finite Markov chain with k possible states and stationary transition distributions.

Example 3.10.5 A Transition Matrix for the Number of Occupied Telephone Lines. Suppose that in the example involving the office with five telephone lines, the numbers of lines being used at times 1, 2, . . . form a Markov chain with stationary transition distributions. This chain has six possible states 0, 1, . . . , 5, where i is the state in which exactly i lines are being used at a given time ($i = 0, 1, \ldots, 5$). Suppose that the transition matrix $\boldsymbol{P}$ is as follows:

$$\boldsymbol{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.1 & 0.4 & 0.2 & 0.1 & 0.1 & 0.1 \\ 0.2 & 0.3 & 0.2 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.3 & 0.2 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.2 & 0.3 & 0.2 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.2 & 0.3 & 0.2 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.4 & 0.2 \end{bmatrix} \end{array}. \tag{3.10.6}$$

(a) Assuming that all five lines are in use at a certain observation time, we shall determine the probability that exactly four lines will be in use at the next observation time. (b) Assuming that no lines are in use at a certain time, we shall determine the probability that at least one line will be in use at the next observation time.

(a) This probability is the element in the matrix $\boldsymbol{P}$ in the row corresponding to the state 5 and the column corresponding to the state 4. Its value is seen to be 0.4.

(b) If no lines are in use at a certain time, then the element in the upper left corner of the matrix $\boldsymbol{P}$ gives the probability that no lines will be in use at the next observation time. Its value is seen to be 0.1. Therefore, the probability that at least one line will be in use at the next observation time is $1 - 0.1 = 0.9$. ◀

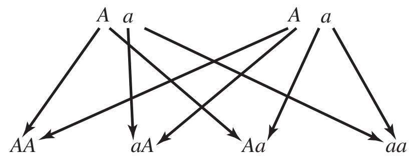


Figure 3.28 The generation following $\{Aa, Aa\}$.

Example 3.10.6

Plant Breeding Experiment. A botanist is studying a certain variety of plant that is monoecious (has male and female organs in separate flowers on a single plant). She begins with two plants I and II and cross-pollinates them by crossing male I with female II and female I with male II to produce two offspring for the next generation. The original plants are destroyed and the process is repeated as soon as the new generation of two plants is mature. Several replications of the study are run simultaneously. The botanist might be interested in the proportion of plants in any generation that have each of several possible genotypes for a particular gene. (See Example 1.6.4 on page 23.) Suppose that the gene has two alleles, A and a. The genotype of an individual will be one of the three combinations AA, Aa, or aa. When a new individual is born, it gets one of the two alleles (with probability 1/2 each) from one of the parents, and it independently gets one of the two alleles from the other parent. The two offspring get their genotypes independently of each other. For example, if the parents have genotypes AA and Aa, then an offspring will get A for sure from the first parent and will get either A or a from the second parent with probability 1/2 each. Let the states of this population be the set of genotypes of the two members of the current population. We will not distinguish the set $\{AA, Aa\}$ from $\{Aa, AA\}$. There are then six states: $\{AA, AA\}$, $\{AA, Aa\}$, $\{AA, aa\}$, $\{Aa, Aa\}$, $\{Aa, aa\}$, and $\{aa, aa\}$. For each state, we can calculate the probability that the next generation will be in each of the six states. For example, if the state is either $\{AA, AA\}$ or $\{aa, aa\}$, the next generation will be in the same state with probability 1. If the state is $\{AA, aa\}$, the next generation will be in state $\{Aa, Aa\}$ with probability 1. The other three states have more complicated transitions.

If the current state is $\{Aa, Aa\}$, then all six states are possible for the next generation. In order to compute the transition distribution, it helps to first compute the probability that a given offspring will have each of the three genotypes. Figure 3.28 illustrates the possible offspring in this state. Each arrow going down in Fig. 3.28 is a possible inheritance of an allele, and each combination of arrows terminating in a genotype has probability 1/4. It follows that the probability of AA and aa are both 1/4, while the probability of Aa is 1/2, because two different combinations of arrows lead to this offspring. In order for the next state to be $\{AA, AA\}$, both offspring must be AA independently, so the probability of this transition is 1/16. The same argument implies that the probability of a transition to $\{aa, aa\}$ is 1/16. A transition to $\{AA, Aa\}$ requires one offspring to be AA (probability 1/4) and the other to be Aa (probabilty 1/2). But the two different genotypes could occur in either order, so the whole probability of such a transition is $2 \times (1/4) \times (1/2) = 1/4$. A similar argument shows that a transition to $\{Aa, aa\}$ also has probability 1/4. A transition to $\{AA, aa\}$ requires one offspring to be AA (probability 1/4) and the other to be aa (probability 1/4). Once again, these can occur in two orders, so the whole probability is $2 \times 1/4 \times 1/4 = 1/8$. By subtraction, the probability of a transition to $\{Aa, Aa\}$ must be $1 - 1/16 - 1/16 - 1/4 - 1/4 - 1/8 = 1/4$. Here is the entire transition matrix, which can be verified in a manner similar to what has just been done:

$$\begin{array}{c} \\ \{AA, AA\} \\ \{AA, Aa\} \\ \{AA, aa\} \\ \{Aa, Aa\} \\ \{Aa, aa\} \\ \{aa, aa\} \end{array}
\begin{array}{c} \begin{array}{cccccc} \{AA, AA\} & \{AA, Aa\} & \{AA, aa\} & \{Aa, Aa\} & \{Aa, aa\} & \{aa, aa\} \end{array} \\
\left[\begin{array}{cccccc}
1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.2500 & 0.5000 & 0.0000 & 0.2500 & 0.0000 & 0.0000 \\
0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 \\
0.0625 & 0.2500 & 0.1250 & 0.2500 & 0.2500 & 0.0625 \\
0.0000 & 0.0000 & 0.0000 & 0.2500 & 0.5000 & 0.2500 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000
\end{array}\right] \end{array}.$$

◄

The Transition Matrix for Several Steps

Example 3.10.7 Single Server Queue. A manager usually checks the server at her store every 5 minutes to see whether the server is busy or not. She models the state of the server (1 = busy or 2 = not busy) as a Markov chain with two possible states and stationary transition distributions given by the following matrix:

$$\boldsymbol{P} = \begin{array}{c} \\ \text{Busy} \\ \text{Not busy} \end{array} \begin{array}{c} \begin{array}{cc} \text{Busy} & \text{Not busy} \end{array} \\ \left[\begin{array}{cc} 0.9 & 0.1 \\ 0.6 & 0.4 \end{array}\right] \end{array}.$$

The manager realizes that, later in the day, she will have to be away for 10 minutes and will miss one server check. She wants to compute the conditional distribution of the state two time periods in the future given each of the possible states. She reasons as follows: If $X_n = 1$ for example, then the state will have to be either 1 or 2 at time $n + 1$ even though she does not care now about the state at time $n + 1$. But, if she computes the joint conditional distribution of X_{n+1} and X_{n+2} given $X_n = 1$, she can sum over the possible values of X_{n+1} to get the conditional distribution of X_{n+2} given $X_n = 1$. In symbols,

$$\begin{aligned}\Pr(X_{n+2} = 1|X_n = 1) = {} & \Pr(X_{n+1} = 1, X_{n+2} = 1|X_n = 1) \\ & + \Pr(X_{n+1} = 2, X_{n+2} = 1|X_n = 1).\end{aligned}$$

By the second part of Theorem 3.10.1,

$$\begin{aligned}\Pr(X_{n+1} = 1, X_{n+2} = 1|X_n = 1) &= \Pr(X_{n+1} = 1|X_n = 1)\Pr(X_{n+2} = 1|X_{n+1} = 1) \\ &= 0.9 \times 0.9 = 0.81.\end{aligned}$$

Similarly,

$$\begin{aligned}\Pr(X_{n+1} = 2, X_{n+2} = 1|X_n = 1) &= \Pr(X_{n+1} = 2|X_n = 1)\Pr(X_{n+2} = 1|X_{n+1} = 2) \\ &= 0.1 \times 0.6 = 0.06.\end{aligned}$$

It follows that $\Pr(X_{n+2} = 1|X_n = 1) = 0.81 + 0.06 = 0.87$, and hence $\Pr(X_{n+2} = 2|X_n = 1) = 1 - 0.87 = 0.13$. By similar reasoning, if $X_n = 2$,

$$\Pr(X_{n+2} = 1|X_n = 2) = 0.6 \times 0.9 + 0.4 \times 0.6 = 0.78,$$

and $\Pr(X_{n+2} = 2|X_n = 2) = 1 - 0.78 = 0.22$. ◄

Generalizing the calculations in Example 3.10.7 to three or more transitions might seem tedious. However, if one examines the calculations carefully, one sees a pattern

that will allow a compact calculation of transition distributions for several steps. Consider a general Markov chain with k possible states $1, \dots, k$ and the transition matrix $\boldsymbol{P}$ given by Eq. (3.10.5). Assuming that the chain is in state i at a given time n, we shall now determine the probability that the chain will be in state j at time $n+2$. In other words, we shall determine the conditional probability of $X_{n+2} = j$ given $X_n = i$. The notation for this probability is $p_{ij}^{(2)}$.

We argue as the manager did in Example 3.10.7. Let r denote the value of X_{n+1} that is not of primary interest but is helpful to the calculation. Then

$$\begin{aligned}
p_{ij}^{(2)} &= \Pr(X_{n+2} = j \mid X_n = i) \\
&= \sum_{r=1}^{k} \Pr(X_{n+1} = r \quad \text{and} \quad X_{n+2} = j \mid X_n = i) \\
&= \sum_{r=1}^{k} \Pr(X_{n+1} = r \mid X_n = i) \Pr(X_{n+2} = j \mid X_{n+1} = r, X_n = i) \\
&= \sum_{r=1}^{k} \Pr(X_{n+1} = r \mid X_n = i) \Pr(X_{n+2} = j \mid X_{n+1} = r) \\
&= \sum_{r=1}^{k} p_{ir} p_{rj},
\end{aligned}$$

where the third equality follows from Theorem 2.1.3 and the fourth equality follows from the definition of a Markov chain.

The value of $p_{ij}^{(2)}$ can be determined in the following manner: If the transition matrix $\boldsymbol{P}$ is squared, that is, if the matrix $\boldsymbol{P}^2 = \boldsymbol{P}\boldsymbol{P}$ is constructed, then the element in the ith row and the jth column of the matrix $\boldsymbol{P}^2$ will be $\sum_{r=1}^{k} p_{ir} p_{rj}$. Therefore, $p_{ij}^{(2)}$ will be the element in the ith row and the jth column of $\boldsymbol{P}^2$.

By a similar argument, the probability that the chain will move from the state i to the state j in three steps, or $p_{ij}^{(3)} = \Pr(X_{n+3} = j \mid X_n = i)$, can be found by constructing the matrix $\boldsymbol{P}^3 = \boldsymbol{P}^2\boldsymbol{P}$. Then the probability $p_{ij}^{(3)}$ will be the element in the ith row and the jth column of the matrix $\boldsymbol{P}^3$.

In general, we have the following result.

Theorem 3.10.2 Multiple Step Transitions. Let $\boldsymbol{P}$ be the transition matrix of a finite Markov chain with stationary transition distributions. For each $m = 2, 3, \dots$, the mth power $\boldsymbol{P}^m$ of the matrix $\boldsymbol{P}$ has in row i and column j the probability $p_{ij}^{(m)}$ that the chain will move from state i to state j in m steps. ■

Definition 3.10.6 Multiple Step Transition Matrix. Under the conditions of Theorem 3.10.2, the matrix $\boldsymbol{P}^m$ is called the *m-step transition matrix* of the Markov chain.

In summary, the ith row of the m-step transition matrix gives the conditional distribution of X_{n+m} given $X_n = i$ for all $i = 1, \dots, k$ and all $n, m = 1, 2, \dots$.

Example 3.10.8 The Two-Step and Three-Step Transition Matrices for the Number of Occupied Telephone Lines. Consider again the transition matrix $\boldsymbol{P}$ given by Eq. (3.10.6) for the Markov chain based on five telephone lines. We shall assume first that i lines are in use at a

certain time, and we shall determine the probability that exactly j lines will be in use two time periods later.

If we multiply the matrix $\boldsymbol{P}$ by itself, we obtain the following two-step transition matrix:

$$\boldsymbol{P}^2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.14 & 0.23 & 0.20 & 0.15 & 0.16 & 0.12 \\ 0.13 & 0.24 & 0.20 & 0.15 & 0.16 & 0.12 \\ 0.12 & 0.20 & 0.21 & 0.18 & 0.17 & 0.12 \\ 0.11 & 0.17 & 0.19 & 0.20 & 0.20 & 0.13 \\ 0.11 & 0.16 & 0.16 & 0.18 & 0.24 & 0.15 \\ 0.11 & 0.16 & 0.15 & 0.17 & 0.25 & 0.16 \end{bmatrix} \end{array}. \tag{3.10.7}$$

From this matrix we can find any two-step transition probability for the chain, such as the following:

i. If two lines are in use at a certain time, then the probability that four lines will be in use two time periods later is 0.17.
ii. If three lines are in use at a certain time, then the probability that three lines will again be in use two time periods later is 0.20.

We shall now assume that i lines are in use at a certain time, and we shall determine the probability that exactly j lines will be in use three time periods later.

If we construct the matrix $\boldsymbol{P}^3 = \boldsymbol{P}^2\boldsymbol{P}$, we obtain the following three-step transition matrix:

$$\boldsymbol{P}^3 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.123 & 0.208 & 0.192 & 0.166 & 0.183 & 0.128 \\ 0.124 & 0.207 & 0.192 & 0.166 & 0.183 & 0.128 \\ 0.120 & 0.197 & 0.192 & 0.174 & 0.188 & 0.129 \\ 0.117 & 0.186 & 0.186 & 0.179 & 0.199 & 0.133 \\ 0.116 & 0.181 & 0.177 & 0.176 & 0.211 & 0.139 \\ 0.116 & 0.180 & 0.174 & 0.174 & 0.215 & 0.141 \end{bmatrix} \end{array}. \tag{3.10.8}$$

From this matrix we can find any three-step transition probability for the chain, such as the following:

i. If all five lines are in use at a certain time, then the probability that no lines will be in use three time periods later is 0.116.
ii. If one line is in use at a certain time, then the probability that exactly one line will again be in use three time periods later is 0.207. ◄

Example 3.10.9

Plant Breeding Experiment. In Example 3.10.6, the transition matrix has many zeros, since many of the transitions will not occur. However, if we are willing to wait two steps, we will find that the only transitions that cannot occur in two steps are those from the first state to anything else and those from the last state to anything else.

Here is the two-step transition matrix:

$$
\begin{array}{c} \\ \{AA, AA\} \\ \{AA, Aa\} \\ \{AA, aa\} \\ \{Aa, Aa\} \\ \{Aa, aa\} \\ \{aa, aa\} \end{array}
\begin{array}{c}
\begin{array}{cccccc} \{AA, AA\} & \{AA, Aa\} & \{AA, aa\} & \{Aa, Aa\} & \{Aa, aa\} & \{aa, aa\} \end{array} \\
\left[\begin{array}{cccccc}
1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
0.3906 & 0.3125 & 0.0313 & 0.1875 & 0.0625 & 0.0156 \\
0.0625 & 0.2500 & 0.1250 & 0.2500 & 0.2500 & 0.0625 \\
0.1406 & 0.1875 & 0.0313 & 0.3125 & 0.1875 & 0.1406 \\
0.0156 & 0.0625 & 0.0313 & 0.1875 & 0.3125 & 0.3906 \\
0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000
\end{array}\right]
\end{array}.
$$

Indeed, if we look at the three-step or the four-step or the general m-step transition matrix, the first and last rows will always be the same. ◄

The first and last states in Example 3.10.9 have the property that, once the chain gets into one of those states, it can't get out. Such states occur in many Markov chains and have a special name.

Definition 3.10.7 **Absorbing State.** In a Markov chain, if $p_{ii} = 1$ for some state i, then that state is called an *absorbing state*.

In Example 3.10.9, there is positive probability of getting into each absorbing state in two steps no matter where the chain starts. Hence, the probability is 1 that the chain will eventually be absorbed into one of the absorbing states if it is allowed to run long enough.

The Initial Distribution

Example 3.10.10 **Single Server Queue.** The manager in Example 3.10.7 enters the store thinking that the probability is 0.3 that the server will be busy the first time that she checks. Hence, the probability is 0.7 that the server will be not busy. These values specify the marginal distribution of the state at time 1, X_1. We can represent this distribution by the vector $\boldsymbol{v} = (0.3, 0.7)$ that gives the probabilities of the two states at time 1 in the same order that they appear in the transition matrix. ◄

The vector giving the marginal distribution of X_1 in Example 3.10.10 has a special name.

Definition 3.10.8 **Probability Vector/Initial Distribution.** A vector consisting of nonnegative numbers that add to 1 is called a *probability vector*. A probability vector whose coordinates specify the probabilities that a Markov chain will be in each of its states at time 1 is called the *initial distribution* of the chain or the *intial probability vector*.

For Example 3.10.2, the initial distribution was given in Exercise 4 in Sec. 2.1 as $\boldsymbol{v} = (0.5, 0.5)$.

The initial distribution and the transition matrix together determine the entire joint distribution of the Markov chain. Indeed, Theorem 3.10.1 shows how to construct the joint distribution of the chain from the initial probability vector and the transition matrix. Letting $\boldsymbol{v} = (v_1, \ldots, v_k)$ denote the initial distribution, Eq. (3.10.1) can be rewritten as

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = v_{x_1} p_{x_1 x_2} \cdots p_{x_{n-1} x_n}. \tag{3.10.9}$$

The marginal distributions of states at times later than 1 can be found from the joint distribution.

Theorem 3.10.3 Marginal Distributions at Times Other Than 1. Consider a finite Markov chain with stationary transition distributions having initial distribution $\boldsymbol{v}$ and transition matrix $\boldsymbol{P}$. The marginal distribution of X_n, the state at time n, is given by the probability vector $\boldsymbol{v}\boldsymbol{P}^{n-1}$.

Proof The marginal distribution of X_n can be found from Eq. (3.10.9) by summing over the possible values of $x_1, \ldots, x_{n-1}$. That is,

$$\Pr(X_n = x_n) = \sum_{x_{n-1}=1}^{k} \cdots \sum_{x_2=1}^{k} \sum_{x_1=1}^{k} v_{x_1} p_{x_1 x_2} p_{x_2 x_3} \cdots p_{x_{n-1} x_n}. \tag{3.10.10}$$

The innermost sum in Eq. (3.10.10) for $x_1 = 1, \ldots, k$ involves only the first two factors $v_{x_1} p_{x_1 x_2}$ and produces the x_2 coordinate of $\boldsymbol{vP}$. Similarly, the next innermost sum over $x_2 = 1, \ldots, k$ involves only the x_2 coordinate of $\boldsymbol{vP}$ and $p_{x_2 x_3}$ and produces the x_3 coordinate of $\boldsymbol{vPP} = \boldsymbol{vP}^2$. Proceeding in this way through all $n - 1$ summations produces the x_n coordinate of $\boldsymbol{vP}^{n-1}$. ■

Example 3.10.11 Probabilities for the Number of Occupied Telephone Lines. Consider again the office with five telephone lines and the Markov chain for which the transition matrix $\boldsymbol{P}$ is given by Eq. (3.10.6). Suppose that at the beginning of the observation process at time $n = 1$, the probability that no lines will be in use is 0.5, the probability that one line will be in use is 0.3, and the probability that two lines will be in use is 0.2. Then the initial probability vector is $\boldsymbol{v} = (0.5, 0.3, 0.2, 0, 0, 0)$. We shall first determine the distribution of the number of lines in use at time 2, one period later.

By an elementary computation it will be found that

$$\boldsymbol{vP} = (0.13, 0.33, 0.22, 0.12, 0.10, 0.10).$$

Since the first component of this probability vector is 0.13, the probability that no lines will be in use at time 2 is 0.13; since the second component is 0.33, the probability that exactly one line will be in use at time 2 is 0.33; and so on.

Next, we shall determine the distribution of the number of lines that will be in use at time 3.

By use of Eq. (3.10.7), it can be found that

$$\boldsymbol{vP}^2 = (0.133, 0.227, 0.202, 0.156, 0.162, 0.120).$$

Since the first component of this probability vector is 0.133, the probability that no lines will be in use at time 3 is 0.133; since the second component is 0.227, the probability that exactly one line will be in use at time 3 is 0.227; and so on. ◄

Stationary Distributions

Example 3.10.12 A Special Initial Distribution for Telephone Lines. Suppose that the initial distribution for the number of occupied telephone lines is

$$\boldsymbol{v} = (0.119, 0.193, 0.186, 0.173, 0.196, 0.133).$$

It can be shown, by matrix multiplication, that $\boldsymbol{vP} = \boldsymbol{v}$. This means that if $\boldsymbol{v}$ is the initial distribution, then it is also the distribution after one transition. Hence, it will also be the distribution after two or more transitions as well. ◄

Definition 3.10.9 Stationary Distribution. Let $\boldsymbol{P}$ be the transition matrix for a Markov chain. A probability vector $\boldsymbol{v}$ that satisfies $\boldsymbol{vP} = \boldsymbol{v}$ is called a *stationary distribution* for the Markov chain.

The initial distribution in Example 3.10.12 is a stationary distribution for the telephone lines Markov chain. If the chain starts in this distribution, the distribution stays the same at all times. Every finite Markov chain with stationary transition distributions has at least one stationary distribution. Some chains have a unique stationary distribution.

Note: A Stationary Distribution Does Not Mean That the Chain is Not Moving. It is important to note that $\boldsymbol{vP}^n$ gives the probabilities that the chain is in each of its states after n transitions, calculated before the initial state of the chain or any transitions are observed. These are different from the probabilities of being in the various states after observing the initial state or after observing any of the intervening transitions. In addition, a stationary distribution does not imply that the Markov chain is staying put. If a Markov chain starts in a stationary distribution, then for each state i, the probability that the chain is in state i after n transitions is the same as the probability that it is state i at the start. But the Markov chain can still move around from one state to the next at each transition. The one case in which a Markov chain does stay put is after it moves into an absorbing state. A distribution that is concentrated solely on absorbing states will necessarily be stationary because the Markov chain will never move if it starts in such a distribution. In such cases, all of the uncertainty surrounds the initial state, which will also be the state after every transition.

Example 3.10.13 Stationary Distributions for the Plant Breeding Experiment. Consider again the experiment described in Example 3.10.6. The first and sixth states, $\{AA, AA\}$ and $\{aa, aa\}$, respectively, are absorbing states. It is easy to see that every initial distribution of the form $\boldsymbol{v} = (p, 0, 0, 0, 0, 1 - p)$ for $0 \le p \le 1$ has the property that $\boldsymbol{vP} = \boldsymbol{v}$. Suppose that the chain is in state 1 with probability p and in state 6 with probability $1 - p$ at time 1. Because these two states are absorbing states, the chain will never move and the event $X_1 = 1$ is the same as the event that $X_n = 1$ for all n. Similarly, $X_1 = 6$ is the same as $X_n = 6$. So, thinking ahead to where the chain is likely to be after n transitions, we would also say that it will be in state 1 with probability p and in state 6 with probability $1 - p$. ◀

Method for Finding Stationary Distributions We can rewrite the equation $\boldsymbol{vP} = \boldsymbol{v}$ that defines stationary distributions as $\boldsymbol{v}[\boldsymbol{P} - \boldsymbol{I}] = \boldsymbol{0}$, where $\boldsymbol{I}$ is a $k \times k$ identity matrix and $\boldsymbol{0}$ is a k-dimensional vector of all zeros. Unfortunately, this system of equations has lots of solutions even if there is a unique stationary distribution. The reason is that whenever $\boldsymbol{v}$ solves the system, so does $c\boldsymbol{v}$ for all real c (including $c = 0$). Even though the system has k equations for k variables, there is at least one redundant equation. However, there is also one missing equation. We need to require that the solution vector $\boldsymbol{v}$ has coordinates that sum to 1. We can fix both of these problems by replacing one of the equations in the original system by the equation that says that the coordinates of $\boldsymbol{v}$ sum to 1.

To be specific, define the matrix $\boldsymbol{G}$ to be $\boldsymbol{P} - \boldsymbol{I}$ with its last column replaced by a column of all ones. Then, solve the equation

$$\boldsymbol{v}\boldsymbol{G} = (0, \ldots, 0, 1). \tag{3.10.11}$$

If there is a unique stationary distribution, we will find it by solving (3.10.11). In this case, the matrix $\boldsymbol{G}$ will have an inverse $\boldsymbol{G}^{-1}$ that satisfies

$$\boldsymbol{G}\boldsymbol{G}^{-1} = \boldsymbol{G}^{-1}\boldsymbol{G} = \boldsymbol{I}.$$

The solution of (3.10.11) will then be

$$\boldsymbol{v} = (0, \ldots, 0, 1)\boldsymbol{G}^{-1},$$

which is easily seen to be the bottom row of the matrix $\boldsymbol{G}^{-1}$. This was the method used to find the stationary distribution in Example 3.10.12. If the Markov chain has multiple stationary distributions, then the matrix $\boldsymbol{G}$ will be singular, and this method will not find any of the stationary distributions. That is what would happen in Example 3.10.13 if one were to apply the method.

Example 3.10.14 Stationary Distribution for Toothpaste Shopping. Consider the transition matrix $\boldsymbol{P}$ given in Example 3.10.4. We can construct the matrix $\boldsymbol{G}$ as follows:

$$\boldsymbol{P} - \boldsymbol{I} = \begin{bmatrix} -\frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{2}{3} \end{bmatrix}; \quad \text{hence } \boldsymbol{G} = \begin{bmatrix} -\frac{2}{3} & 1 \\ \frac{2}{3} & 1 \end{bmatrix}.$$

The inverse of $\boldsymbol{G}$ is

$$\boldsymbol{G}^{-1} = \begin{bmatrix} -\frac{3}{4} & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

We now see that the stationary distribution is the bottom row of G^{-1}, $\boldsymbol{v} = (1/2, 1/2)$. ◀

There is a special case in which it is known that a unique stationary distribution exists and it has special properties.

Theorem 3.10.4 If there exists m such that every element of $\boldsymbol{P}^m$ is strictly positive, then

- the Markov chain has a unique stationary distribution $\boldsymbol{v}$,
- $\lim_{n\to\infty} \boldsymbol{P}^n$ is a matrix with all rows equal to $\boldsymbol{v}$, and
- no matter with what distribution the Markov chain starts, its distribution after n steps converges to $\boldsymbol{v}$ as $n \to \infty$. ■

We shall not prove Theorem 3.10.4, although some evidence for the second claim can be seen in Eq. (3.10.8), where the six rows of $\boldsymbol{P}^3$ are much more alike than the rows of $\boldsymbol{P}$ and they are very similar to the stationary distribution given in Example 3.10.12. The third claim in Theorem 3.10.4 actually follows easily from the second claim. In Sec. 12.5, we shall introduce a method that makes use of the third claim in Theorem 3.10.4 in order to approximate distributions of random variables when those distributions are difficult to calculate exactly.

The transition matrices in Examples 3.10.2, 3.10.5, and 3.10.7 satisfy the conditions of Theorem 3.10.4. The following example has a unique stationary distribution but does not satisfy the conditions of Theorem 3.10.4.

Example 3.10.15 Alternating Chain. Let the transition matrix for a two-state Markov chain be

$$\boldsymbol{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The matrix $\boldsymbol{G}$ is easy to construct and invert, and we find that the unique stationary distribution is $\boldsymbol{v} = (0.5, 0.5)$. However, as m increases, $\boldsymbol{P}^m$ alternates between $\boldsymbol{P}$ and the 2×2 identity matrix. It does not converge and never does it have all elements strictly positive. If the initial distribution is (v_1, v_2), the distribution after n steps alternates between (v_1, v_2) and (v_2, v_1). ◂

Another example that fails to satisfy the conditions of Theorem 3.10.4 is the gambler's ruin problem from Sec. 2.4.

Example 3.10.16 Gambler's Ruin. In Sec. 2.4, we described the gambler's ruin problem, in which a gambler wins one dollar with probability p and loses one dollar with probability $1 - p$ on each play of a game. The sequence of amounts held by the gambler through the course of those plays forms a Markov chain with two absorbing states, namely, 0 and k. There are $k - 1$ other states, namely, $1, \dots, k - 1$. (This notation violates our use of k to stand for the number of states, which is $k + 1$ in this example. We felt this was less confusing than switching from the original notation of Sec. 2.4.) The transition matrix has first and last row being $(1, 0, \dots, 0)$ and $(0, \dots, 1)$, respectively. The ith row (for $i = 1, \dots, k - 1$) has 0 everywhere except in coordinate $i - 1$ where it has $1 - p$ and in coordinate $i + 1$ where it has p. Unlike Example 3.10.15, this time the sequence of matrices $\boldsymbol{P}^m$ converges but there is no unique stationary distribution. The limit of $\boldsymbol{P}^m$ has as its last column the numbers $a_0, \dots, a_k$, where a_i is the probability that the fortune of a gambler who starts with i dollars reaches k dollars before it reaches 0 dollars. The first column of the limit has the numbers $1 - a_0, \dots, 1 - a_k$ and the rest of the limit matrix is all zeros. The stationary distributions have the same form as those in Example 3.10.13, namely, all probability is in the absorbing states. ◂

Summary

A Markov chain is a stochastic process, a sequence of random variables giving the states of the process, in which the conditional distribution of the state at the next time given all of the past states depends on the past states only through the most recent state. For Markov chains with finitely many states and stationary transition distributions, the transitions over time can be described by a matrix giving the probabilities of transition from the state indexing the row to the state indexing the column (the transition matrix $\boldsymbol{P}$). The initial probability vector $\boldsymbol{v}$ gives the distribution of the state at time 1. The transition matrix and initial probability vector together allow calculation of all probabilities associated with the Markov chain. In particular, $\boldsymbol{P}^n$ gives the probabilities of transitions over n time periods, and $\boldsymbol{v}\boldsymbol{P}^n$ gives the distribution of the state at time $n + 1$. A stationary distribution is a probability vector $\boldsymbol{v}$ such that $\boldsymbol{v}\boldsymbol{P} = \boldsymbol{v}$. Every finite Markov chain with stationary transition distributions has at least one stationary distribution. For many Markov chains, there is a unique stationary distribution and the distribution of the chain after n transitions converges to the stationary distribution as n goes to ∞.

Exercises

1. Consider the Markov chain in Example 3.10.2 with initial probability vector $\boldsymbol{v} = (1/2, 1/2)$.

a. Find the probability vector specifying the probabilities of the states at time $n = 2$.

b. Find the two-step transition matrix.

2. Suppose that the weather can be only sunny or cloudy and the weather conditions on successive mornings form a Markov chain with stationary transition probabilities. Suppose also that the transition matrix is as follows:

	Sunny	Cloudy
Sunny	0.7	0.3
Cloudy	0.6	0.4

a. If it is cloudy on a given day, what is the probability that it will also be cloudy the next day?

b. If it is sunny on a given day, what is the probability that it will be sunny on the next two days?

c. If it is cloudy on a given day, what is the probability that it will be sunny on at least one of the next three days?

3. Consider again the Markov chain described in Exercise 2.

a. If it is sunny on a certain Wednesday, what is the probability that it will be sunny on the following Saturday?

b. If it is cloudy on a certain Wednesday, what is the probability that it will be sunny on the following Saturday?

4. Consider again the conditions of Exercises 2 and 3.

a. If it is sunny on a certain Wednesday, what is the probability that it will be sunny on both the following Saturday and Sunday?

b. If it is cloudy on a certain Wednesday, what is the probability that it will be sunny on both the following Saturday and Sunday?

5. Consider again the Markov chain described in Exercise 2. Suppose that the probability that it will be sunny on a certain Wednesday is 0.2 and the probability that it will be cloudy is 0.8.

a. Determine the probability that it will be cloudy on the next day, Thursday.

b. Determine the probability that it will be cloudy on Friday.

c. Determine the probability that it will be cloudy on Saturday.

6. Suppose that a student will be either on time or late for a particular class and that the events that he is on time or late for the class on successive days form a Markov chain with stationary transition probabilities. Suppose also that if he is late on a given day, then the probability that he will be on time the next day is 0.8. Furthermore, if he is on time on a given day, then the probability that he will be late the next day is 0.5.

a. If the student is late on a certain day, what is the probability that he will be on time on each of the next three days?

b. If the student is on time on a given day, what is the probability that he will be late on each of the next three days?

7. Consider again the Markov chain described in Exercise 6.

a. If the student is late on the first day of class, what is the probability that he will be on time on the fourth day of class?

b. If the student is on time on the first day of class, what is the probability that he will be on time on the fourth day of class?

8. Consider again the conditions of Exercises 6 and 7. Suppose that the probability that the student will be late on the first day of class is 0.7 and that the probability that he will be on time is 0.3.

a. Determine the probability that he will be late on the second day of class.

b. Determine the probability that he will be on time on the fourth day of class.

9. Suppose that a Markov chain has four states 1, 2, 3, 4 and stationary transition probabilities as specified by the following transition matrix:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 1/4 & 1/4 & 0 & 1/2 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix} \end{array}.$$

a. If the chain is in state 3 at a given time n, what is the probability that it will be in state 2 at time $n+2$?

b. If the chain is in state 1 at a given time n, what is the probability that it will be in state 3 at time $n+3$?

10. Let X_1 denote the initial state at time 1 of the Markov chain for which the transition matrix is as specified in Exercise 5, and suppose that the initial probabilities are as follows:

$$\Pr(X_1 = 1) = 1/8,\ \Pr(X_1 = 2) = 1/4,$$
$$\Pr(X_1 = 3) = 3/8,\ \Pr(X_1 = 4) = 1/4.$$

Determine the probabilities that the chain will be in states 1, 2, 3, and 4 at time n for each of the following values of n: **(a)** $n = 2$; **(b)** $n = 3$; **(c)** $n = 4$.

11. Each time that a shopper purchases a tube of toothpaste, she chooses either brand A or brand B. Suppose that the probability is 1/3 that she will choose the same brand

chosen on her previous purchase, and the probability is 2/3 that she will switch brands.

a. If her first purchase is brand A, what is the probability that her fifth purchase will be brand B?

b. If her first purchase is brand B, what is the probability that her fifth purchase will be brand B?

12. Suppose that three boys A, B, and C are throwing a ball from one to another. Whenever A has the ball, he throws it to B with a probability of 0.2 and to C with a probability of 0.8. Whenever B has the ball, he throws it to A with a probability of 0.6 and to C with a probability of 0.4. Whenever C has the ball, he is equally likely to throw it to either A or B.

a. Consider this process to be a Markov chain and construct the transition matrix.

b. If each of the three boys is equally likely to have the ball at a certain time n, which boy is most likely to have the ball at time $n+2$?

13. Suppose that a coin is tossed repeatedly in such a way that heads and tails are equally likely to appear on any given toss and that all tosses are independent, with the following exception: Whenever either three heads or three tails have been obtained on three successive tosses, then the outcome of the next toss is always of the opposite type. At time n ($n \geq 3$), let the state of this process be specified by the outcomes on tosses $n-2$, $n-1$, and n. Show that this process is a Markov chain with stationary transition probabilities and construct the transition matrix.

14. There are two boxes A and B, each containing red and green balls. Suppose that box A contains one red ball and two green balls and box B contains eight red balls and two green balls. Consider the following process: One ball is selected at random from box A, and one ball is selected at random from box B. The ball selected from box A is then placed in box B and the ball selected from box B is placed in box A. These operations are then repeated indefinitely. Show that the numbers of red balls in box A form a Markov chain with stationary transition probabilities, and construct the transition matrix of the Markov chain.

15. Verify the rows of the transition matrix in Example 3.10.6 that correspond to current states $\{AA, Aa\}$ and $\{Aa, aa\}$.

16. Let the initial probability vector in Example 3.10.6 be $\boldsymbol{v} = (1/16, 1/4, 1/8, 1/4, 1/4, 1/16)$. Find the probabilities of the six states after one generation.

17. Return to Example 3.10.6. Assume that the state at time $n-1$ is $\{Aa, aa\}$.

a. Suppose that we learn that X_{n+1} is $\{AA, aa\}$. Find the conditional distribution of X_n. (That is, find all the probabilities for the possible states at time n given that the state at time $n+1$ is $\{AA, aa\}$.)

b. Suppose that we learn that X_{n+1} is $\{aa, aa\}$. Find the conditional distribution of X_n.

18. Return to Example 3.10.13. Prove that the stationary distributions described there are the only stationary distributions for that Markov chain.

19. Find the unique stationary distribution for the Markov chain in Exercise 2.

20. The unique stationary distribution in Exercise 9 is $\mathbf{v} = (0, 1, 0, 0)$. This is an instance of the following general result: Suppose that a Markov chain has exactly one absorbing state. Suppose further that, for each non-absorbing state k, there is n such that the probability is positive of moving from state k to the absorbing state in n steps. Then the unique stationary distribution has probability 1 in the absorbing state. Prove this result.

3.11 Supplementary Exercises

1. Suppose that X and Y are independent random variables, that X has the uniform distribution on the integers 1, 2, 3, 4, 5 (discrete), and that Y has the uniform distribution on the interval [0, 5] (continuous). Let Z be a random variable such that $Z = X$ with probability 1/2 and $Z = Y$ with probability 1/2. Sketch the c.d.f. of Z.

2. Suppose that X and Y are independent random variables. Suppose that X has a discrete distribution concentrated on finitely many distinct values with p.f. f_1. Suppose that Y has a continuous distribution with p.d.f. f_2. Let $Z = X + Y$. Show that Z has a continuous distribution and find its p.d.f. *Hint:* First find the conditional p.d.f. of Z given $X = x$.

3. Suppose that the random variable X has the following c.d.f.:

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ \frac{2}{5}x & \text{for } 0 < x \leq 1, \\ \frac{3}{5}x - \frac{1}{5} & \text{for } 1 < x \leq 2, \\ 1 & \text{for } x > 2. \end{cases}$$

Verify that X has a continuous distribution, and determine the p.d.f. of X.

4. Suppose that the random variable X has a continuous distribution with the following p.d.f.:

$$f(x) = \frac{1}{2}e^{-|x|} \quad \text{for } -\infty < x < \infty.$$

Determine the value x_0 such that $F(x_0) = 0.9$, where $F(x)$ is the c.d.f. of X.

5. Suppose that X_1 and X_2 are i.i.d. random variables, and that each has the uniform distribution on the interval [0, 1]. Evaluate $\Pr(X_1^2 + X_2^2 \le 1)$.

6. For each value of $p > 1$, let

$$c(p) = \sum_{x=1}^{\infty} \frac{1}{x^p}.$$

Suppose that the random variable X has a discrete distribution with the following p.f.:

$$f(x) = \frac{1}{c(p)x^p} \quad \text{for } x = 1, 2, \ldots.$$

- **a.** For each fixed positive integer n, determine the probability that X will be divisible by n.
- **b.** Determine the probability that X will be odd.

7. Suppose that X_1 and X_2 are i.i.d. random variables, each of which has the p.f. $f(x)$ specified in Exercise 6. Determine the probability that $X_1 + X_2$ will be even.

8. Suppose that an electronic system comprises four components, and let X_j denote the time until component j fails to operate ($j = 1, 2, 3, 4$). Suppose that X_1, X_2, X_3, and X_4 are i.i.d. random variables, each of which has a continuous distribution with c.d.f. $F(x)$. Suppose that the system will operate as long as both component 1 and at least one of the other three components operate. Determine the c.d.f. of the time until the system fails to operate.

9. Suppose that a box contains a large number of tacks and that the probability X that a particular tack will land with its point up when it is tossed varies from tack to tack in accordance with the following p.d.f.:

$$f(x) = \begin{cases} 2(1-x) & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that a tack is selected at random from the box and that this tack is then tossed three times independently. Determine the probability that the tack will land with its point up on all three tosses.

10. Suppose that the radius X of a circle is a random variable having the following p.d.f.:

$$f(x) = \begin{cases} \frac{1}{8}(3x+1) & \text{for } 0 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the p.d.f. of the area of the circle.

11. Suppose that the random variable X has the following p.d.f.:

$$f(x) = \begin{cases} 2e^{-2x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Construct a random variable $Y = r(X)$ that has the uniform distribution on the interval [0, 5].

12. Suppose that the 12 random variables $X_1, \ldots, X_{12}$ are i.i.d. and each has the uniform distribution on the interval [0, 20]. For $j = 0, 1, \ldots, 19$, let I_j denote the interval $(j, j+1)$. Determine the probability that none of the 20 disjoint intervals I_j will contain more than one of the random variables $X_1, \ldots, X_{12}$.

13. Suppose that the joint distribution of X and Y is uniform over a set A in the xy-plane. For which of the following sets A are X and Y independent?

- **a.** A circle with a radius of 1 and with its center at the origin
- **b.** A circle with a radius of 1 and with its center at the point (3, 5)
- **c.** A square with vertices at the four points (1, 1), (1, −1), (−1, −1), and (−1, 1)
- **d.** A rectangle with vertices at the four points (0, 0), (0, 3), (1, 3), and (1, 0)
- **e.** A square with vertices at the four points (0, 0), (1, 1), (0, 2), and (−1, 1)

14. Suppose that X and Y are independent random variables with the following p.d.f.'s:

$$f_1(x) = \begin{cases} 1 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$f_2(y) = \begin{cases} 8y & \text{for } 0 < y < \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the value of $\Pr(X > Y)$.

15. Suppose that, on a particular day, two persons A and B arrive at a certain store independently of each other. Suppose that A remains in the store for 15 minutes and B remains in the store for 10 minutes. If the time of arrival of each person has the uniform distribution over the hour between 9:00 A.M. and 10:00 A.M., what is the probability that A and B will be in the store at the same time?

16. Suppose that X and Y have the following joint p.d.f.:

$$f(x, y) = \begin{cases} 2(x+y) & \text{for } 0 < x < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** $\Pr(X < 1/2)$; **(b)** the marginal p.d.f. of X; and **(c)** the conditional p.d.f. of Y given that $X = x$.

17. Suppose that X and Y are random variables. The marginal p.d.f. of X is

$$f(x) = \begin{cases} 3x^2 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Also, the conditional p.d.f. of Y given that $X = x$ is

$$g(y|x) = \begin{cases} \frac{3y^2}{x^3} & \text{for } 0 < y < x, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the marginal p.d.f. of Y and **(b)** the conditional p.d.f. of X given that $Y = y$.

18. Suppose that the joint distribution of X and Y is uniform over the region in the xy-plane bounded by the four lines $x = -1$, $x = 1$, $y = x + 1$, and $y = x - 1$. Determine **(a)** $\Pr(XY > 0)$ and **(b)** the conditional p.d.f. of Y given that $X = x$.

19. Suppose that the random variables X, Y, and Z have the following joint p.d.f.:

$$f(x, y, z) = \begin{cases} 6 & \text{for } 0 < x < y < z < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the univariate marginal p.d.f.'s of X, Y, and Z.

20. Suppose that the random variables X, Y, and Z have the following joint p.d.f.:

$$f(x, y, z) = \begin{cases} 2 & \text{for } 0 < x < y < 1 \text{ and } 0 < z < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Evaluate $\Pr(3X > Y | 1 < 4Z < 2)$.

21. Suppose that X and Y are i.i.d. random variables, and that each has the following p.d.f.:

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Also, let $U = X/(X + Y)$ and $V = X + Y$.

a. Determine the joint p.d.f. of U and V.

b. Are U and V independent?

22. Suppose that the random variables X and Y have the following joint p.d.f.:

$$f(x, y) = \begin{cases} 8xy & \text{for } 0 \le x \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Also, let $U = X/Y$ and $V = Y$.

a. Determine the joint p.d.f. of U and V.

b. Are X and Y independent?

c. Are U and V independent?

23. Suppose that $X_1, \dots, X_n$ are i.i.d. random variables, each having the following c.d.f.:

$$F(x) = \begin{cases} 0 & \text{for } x \le 0, \\ 1 - e^{-x} & \text{for } x > 0. \end{cases}$$

Let $Y_1 = \min\{X_1, \dots, X_n\}$ and $Y_n = \max\{X_1, \dots, X_n\}$. Determine the conditional p.d.f. of Y_1 given that $Y_n = y_n$.

24. Suppose that X_1, X_2, and X_3 form a random sample of three observations from a distribution having the following p.d.f.:

$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the p.d.f. of the range of the sample.

25. In this exercise, we shall provide an approximate justification for Eq. (3.6.6). First, remember that if a and b are close together, then

$$\int_a^b r(t)dt \approx (b - a)r\left(\frac{a + b}{2}\right). \tag{3.11.1}$$

Throughout this problem, assume that X and Y have joint p.d.f. f.

a. Use (3.11.1) to approximate $\Pr(y - \epsilon < Y \le y + \epsilon)$.

b. Use (3.11.1) with $r(t) = f(s, t)$ for fixed s to approximate

$$\Pr(X \le x \text{ and } y - \epsilon < Y \le y + \epsilon) = \int_{-\infty}^{x} \int_{y-\epsilon}^{y+\epsilon} f(s, t)\, dt\, ds.$$

c. Show that the ratio of the approximation in part (b) to the approximation in part (a) is $\int_{-\infty}^{x} g_1(s|y)\, ds$.

26. Let X_1, X_2 be two independent random variables each with p.d.f. $f_1(x) = e^{-x}$ for $x > 0$ and $f_1(x) = 0$ for $x \le 0$. Let $Z = X_1 - X_2$ and $W = X_1/X_2$.

a. Find the joint p.d.f. of X_1 and Z.

b. Prove that the conditional p.d.f. of X_1 given $Z = 0$ is

$$g_1(x_1|0) = \begin{cases} 2e^{-2x_1} & \text{for } x_1 > 0, \\ 0 & \text{otherwise.} \end{cases}$$

c. Find the joint p.d.f. of X_1 and W.

d. Prove that the conditional p.d.f. of X_1 given $W = 1$ is

$$h_1(x_1|1) = \begin{cases} 4x_1 e^{-2x_1} & \text{for } x_1 > 0, \\ 0 & \text{otherwise.} \end{cases}$$

e. Notice that $\{Z = 0\} = \{W = 1\}$, but the conditional distribution of X_1 given $Z = 0$ is not the same as the conditional distribution of X_1 given $W = 1$. This discrepancy is known as the *Borel paradox*. In light of the discussion that begins on page 146 about how conditional p.d.f.'s are not like conditioning on events of probability 0, show how "Z very close to 0" is not the same as "W very close to 1." *Hint:* Draw a set of axes for x_1 and x_2, and draw the two sets $\{(x_1, x_2) : |x_1 - x_2| < \epsilon\}$ and $\{(x_1, x_2) : |x_1/x_2 - 1| < \epsilon\}$ and see how much different they are.

27. Three boys A, B, and C are playing table tennis. In each game, two of the boys play against each other and the third boy does not play. The winner of any given game n plays again in game $n + 1$ against the boy who did not play in game n, and the loser of game n does not play in game $n + 1$. The probability that A will beat B in any game that they play against each other is 0.3, the probability that A will beat C is 0.6, and the probability that B will beat C is 0.8. Represent this process as a Markov chain with stationary transition probabilities by defining the possible states and constructing the transition matrix.

28. Consider again the Markov chain described in Exercise 27. **(a)** Determine the probability that the two boys who play against each other in the first game will play against each other again in the fourth game. **(b)** Show that this probability does not depend on which two boys play in the first game.

29. Find the unique stationary distribution for the Markov chain in Exercise 27.

This page intentionally left blank

Chapter 4

EXPECTATION

4.1 The Expectation of a Random Variable

The distribution of a random variable X contains all of the probabilistic information about X. The entire distribution of X, however, is usually too cumbersome for presenting this information. Summaries of the distribution, such as the average value, or expected value, can be useful for giving people an idea of where we expect X to be without trying to describe the entire distribution. The expected value also plays an important role in the approximation methods that arise in Chapter 6.

Expectation for a Discrete Distribution

Example 4.1.1 Fair Price for a Stock. An investor is considering whether or not to invest \$18 per share in a stock for one year. The value of the stock after one year, in dollars, will be $18 + X$, where X is the amount by which the price changes over the year. At present X is unknown, and the investor would like to compute an "average value" for X in order to compare the return she expects from the investment to what she would get by putting the \$18 in the bank at 5% interest. ◄

The idea of finding an average value as in Example 4.1.1 arises in many applications that involve a random variable. One popular choice is what we call the *mean* or *expected value* or *expectation*.

The intuitive idea of the mean of a random variable is that it is the weighted average of the possible values of the random variable with the weights equal to the probabilities.

Example 4.1.2 Stock Price Change. Suppose that the change in price of the stock in Example 4.1.1 is a random variable X that can assume only the four different values -2, 0, 1, and 4, and that $\Pr(X = -2) = 0.1$, $\Pr(X = 0) = 0.4$, $\Pr(X = 1) = 0.3$, and $\Pr(X = 4) = 0.2$. Then the weighted avarage of these values is

$$-2(0.1) + 0(0.4) + 1(0.3) + 4(0.2) = 0.9.$$

The investor now compares this with the interest that would be earned on \$18 at 5% for one year, which is $18 \times 0.05 = 0.9$ dollars. From this point of view, the price of \$18 seems fair. ◄

The calculation in Example 4.1.2 generalizes easily to every random variable that assumes only finitely many values. Possible problems arise with random variables that assume more than finitely many values, especially when the collection of possible values is unbounded.

Definition 4.1.1 Mean of Bounded Discrete Random Variable. Let X be a bounded discrete random variable whose p.f. is f. The *expectation of* X, denoted by $E(X)$, is a number defined as follows:

$$E(X) = \sum_{\text{All } x} x f(x). \tag{4.1.1}$$

The expectation of X is also referred to as the *mean of* X or the *expected value of* X.

In Example 4.1.2, $E(X) = 0.9$. Notice that 0.9 is not one of the possible values of X in that example. This is typically the case with discrete random variables.

Example 4.1.3 Bernoulli Random Variable. Let X have the Bernoulli distribution with parameter p, that is, assume that X takes only the two values 0 and 1 with $\Pr(X = 1) = p$. Then the mean of X is

$$E(X) = 0 \times (1 - p) + 1 \times p = p. \quad ◀$$

If X is unbounded, it might still be possible to define $E(X)$ as the weighted average of its possible values. However, some care is needed.

Definition 4.1.2 Mean of General Discrete Random Variable. Let X be a discrete random variable whose p.f. is f. Suppose that at least one of the following sums is finite:

$$\sum_{\text{Positive } x} x f(x), \qquad \sum_{\text{Negative } x} x f(x). \tag{4.1.2}$$

Then the *mean*, *expectation*, or *expected value* of X is said to *exist* and is defined to be

$$E(X) = \sum_{\text{All } x} x f(x). \tag{4.1.3}$$

If both of the sums in (4.1.2) are infinite, then $E(X)$ *does not exist*.

The reason that the expectation fails to exist if both of the sums in (4.1.2) are infinite is that, in such cases, the sum in (4.1.3) is not well-defined. It is known from calculus that the sum of an infinite series whose positive and negative terms both add to infinity either fails to converge or can be made to converge to many different values by rearranging the terms in different orders. We don't want the meaning of expected value to depend on arbitrary choices about what order to add numbers. If only one of two sums in (4.1.3) is infiinte, then the expected value is also infinite with the same sign as that of the sum that is infinite. If both sums are finite, then the sum in (4.1.3) converges and doesn't depend on the order in which the terms are added.

Example 4.1.4 The Mean of X Does Not Exist. Let X be a random variable whose p.f. is

$$f(x) = \begin{cases} \dfrac{1}{2|x|(|x|+1)} & \text{if } x = \pm 1, \pm 2, \pm 3, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

It can be verified that this function satisfies the conditions required to be a p.f. The two sums in (4.1.2) are

$$\sum_{x=-1}^{-\infty} x \frac{1}{2|x|(|x|+1)} = -\infty \quad \text{and} \quad \sum_{x=1}^{\infty} x \frac{1}{2x(x+1)} = \infty;$$

hence, $E(X)$ does not exist. ◀

Example 4.1.5 An Infinite Mean. Let X be a random variable whose p.f. is

$$f(x) = \begin{cases} \dfrac{1}{x(x+1)} & \text{if } x = 1, 2, 3, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

The sum over negative values in Eq. (4.1.2) is 0, so the mean of X exists and is

$$E(X) = \sum_{x=1}^{\infty} x \frac{1}{x(x+1)} = \infty.$$

We say that the mean of X is *infinite* in this case. ◀

Note: The Expectation of X Depends Only on the Distribution of X. Although $E(X)$ is called the expectation of X, it depends only on the distribution of X. Every two random variables that have the same distribution will have the same expectation even if they have nothing to do with each other. For this reason, we shall often refer to the expectation of a distribution even if we do not have in mind a random variable with that distribution.

Expectation for a Continuous Distribution

The idea of computing a weighted average of the possible values can be generalized to continuous random variables by using integrals instead of sums. The distinction between bounded and unbounded random variables arises in this case for the same reasons.

Definition 4.1.3 Mean of Bounded Continuous Random Variable. Let X be a bounded continuous random variable whose p.d.f. is f. The *expectation of* X, denoted $E(X)$, is defined as follows:

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx. \tag{4.1.4}$$

Once again, the expectation is also called the *mean* or the *expected value*.

Example 4.1.6 Expected Failure Time. An appliance has a maximum lifetime of one year. The time X until it fails is a random variable with a continuous distribution having p.d.f.

$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$E(X) = \int_0^1 x(2x)\, dx = \int_0^1 2x^2\, dx = \frac{2}{3}.$$

We can also say that the expectation of the distribution with p.d.f. f is 2/3. ◀

For general continuous random variables, we modify Definition 4.1.2.

Definition 4.1.4 Mean of General Continuous Random Variable. Let X be a continuous random variable whose p.d.f. is f. Suppose that at least one of the following integrals is finite:

$$\int_0^\infty xf(x)dx, \quad \int_{-\infty}^0 xf(x)dx. \tag{4.1.5}$$

Then the *mean*, *expectation*, or *expected value* of X is said to *exist* and is defined to be

$$E(X) = \int_{-\infty}^\infty xf(x)dx. \tag{4.1.6}$$

If both of the integrals in (4.1.5) are infinite, then $E(X)$ *does not exist*.

Example 4.1.7 Failure after Warranty. A product has a warranty of one year. Let X be the time at which the product fails. Suppose that X has a continuous distribution with the p.d.f.

$$f(x) = \begin{cases} 0 & \text{for } x < 1, \\ \frac{2}{x^3} & \text{for } x \ge 1. \end{cases}$$

The expected time to failure is then

$$E(X) = \int_1^\infty x\frac{2}{x^3}dx = \int_1^\infty \frac{2}{x^2}dx = 2.$$ ◀

Example 4.1.8 A Mean That Does Not Exist. Suppose that a random variable X has a continuous distribution for which the p.d.f. is as follows:

$$f(x) = \frac{1}{\pi(1+x^2)} \quad \text{for } -\infty < x < \infty. \tag{4.1.7}$$

This distribution is called the *Cauchy distribution*. We can verify the fact that $\int_{-\infty}^\infty f(x)\,dx = 1$ by using the following standard result from elementary calculus:

$$\frac{d}{dx}\tan^{-1} x = \frac{1}{1+x^2} \quad \text{for } -\infty < x < \infty.$$

The two integrals in (4.1.5) are

$$\int_0^\infty \frac{x}{\pi(1+x^2)}dx = \infty \quad \text{and} \quad \int_{-\infty}^0 \frac{x}{\pi(1+x^2)}dx = -\infty;$$

hence, the mean of X does not exist. ◀

Interpretation of the Expectation

Relation of the Mean to the Center of Gravity The expectation of a random variable or, equivalently, the mean of its distribution can be regarded as being the center of gravity of that distribution. To illustrate this concept, consider, for example, the p.f. sketched in Fig. 4.1. The x-axis may be regarded as a long weightless rod to which weights are attached. If a weight equal to $f(x_j)$ is attached to this rod at each point x_j, then the rod will be balanced if it is supported at the point $E(X)$.

Now consider the p.d.f. sketched in Fig. 4.2. In this case, the x-axis may be regarded as a long rod over which the mass varies continuously. If the density of

Figure 4.1 The mean of a discrete distribution.

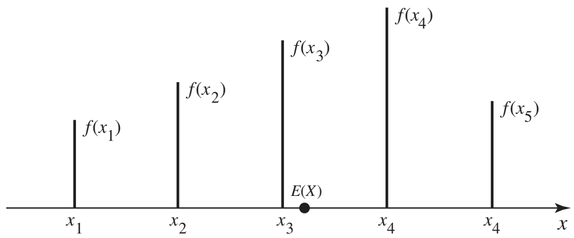
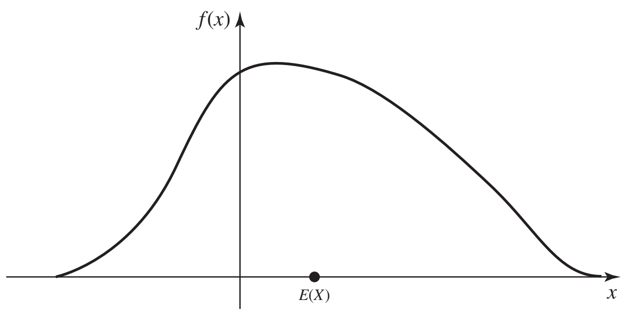


Figure 4.2 The mean of a continuous distribution.



the rod at each point x is equal to $f(x)$, then the center of gravity of the rod will be located at the point $E(X)$, and the rod will be balanced if it is supported at that point.

It can be seen from this discussion that the mean of a distribution can be affected greatly by even a very small change in the amount of probability that is assigned to a large value of x. For example, the mean of the distribution represented by the p.f. in Fig. 4.1 can be moved to any specified point on the x-axis, no matter how far from the origin that point may be, by removing an arbitrarily small but positive amount of probability from one of the points x_j and adding this amount of probability at a point far enough from the origin.

Suppose now that the p.f. or p.d.f. f of some distribution is symmetric with respect to a given point x_0 on the x-axis. In other words, suppose that $f(x_0 + \delta) = f(x_0 - \delta)$ for all values of δ. Also assume that the mean $E(X)$ of this distribution exists. In accordance with the interpretation that the mean is at the center of gravity, it follows that $E(X)$ must be equal to x_0, which is the point of symmetry. The following example emphasizes the fact that it is necessary to make certain that the mean $E(X)$ exists before it can be concluded that $E(X) = x_0$.

Example 4.1.9 The Cauchy Distribution. Consider again the p.d.f. specified by Eq. (4.1.7), which is sketched in Fig. 4.3. This p.d.f. is symmetric with respect to the point $x = 0$. Therefore, if the mean of the Cauchy distribution existed, its value would have to be 0. However, we saw in Example 4.1.8 that the mean of X does not exist.

The reason for the nonexistence of the mean of the Cauchy distribution is as follows: When the curve $y = f(x)$ is sketched as in Fig. 4.3, its tails approach the x-axis rapidly enough to permit the total area under the curve to be equal to 1. On the other hand, if each value of $f(x)$ is multiplied by x and the curve $y = xf(x)$ is sketched, as in Fig. 4.4, the tails of this curve approach the x-axis so slowly that the total area between the x-axis and each part of the curve is infinite. ◀

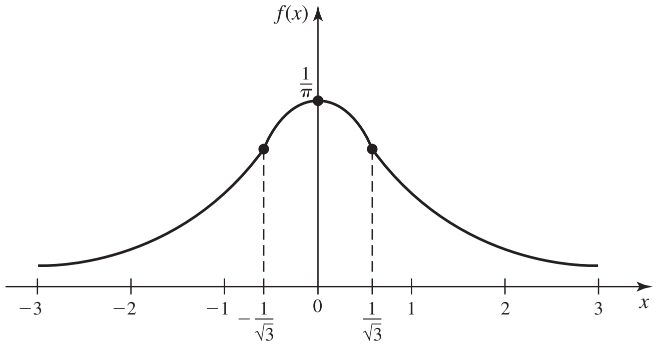


Figure 4.3 The p.d.f. of a Cauchy distribution.

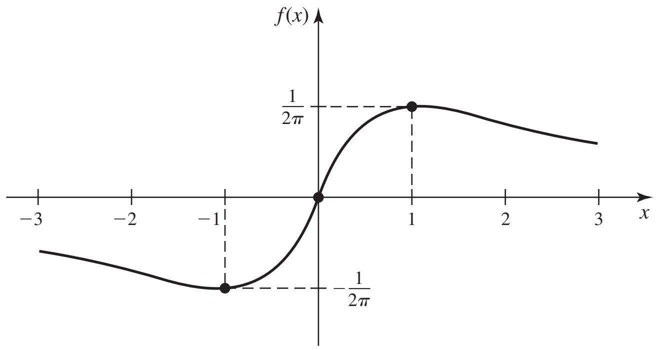


Figure 4.4 The curve $y = xf(x)$ for the Cauchy distribution.

The Expectation of a Function

Example 4.1.10 Failure Rate and Time to Failure. Suppose that appliances manufactured by a particular company fail at a rate of X per year, where X is currently unknown and hence is a random variable. If we are interested in predicting how long such an appliance will last before failure, we might use the mean of $1/X$. How can we calculate the mean of $Y = 1/X$? ◄

Functions of a Single Random Variable If X is a random variable for which the p.d.f. is f, then the expectation of each real-valued function $r(X)$ can be found by applying the definition of expectation to the distribution of $r(X)$ as follows: Let $Y = r(X)$, determine the probability distribution of Y, and then determine $E(Y)$ by applying either Eq. (4.1.1) or Eq. (4.1.4). For example, suppose that Y has a continuous distribution with the p.d.f. g. Then

$$E[r(X)] = E(Y) = \int_{-\infty}^{\infty} yg(y)\,dy, \tag{4.1.8}$$

if the expectation exists.

Example 4.1.11 Failure Rate and Time to Failure. In Example 4.1.10, suppose that the p.d.f. of X is

$$f(x) = \begin{cases} 3x^2 & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let $r(x) = 1/x$. Using the methods of Sec. 3.8, we can find the p.d.f. of $Y = r(X)$ as

$$g(y) = \begin{cases} 3y^{-4} & \text{if } y > 1, \\ 0 & \text{otherwise.} \end{cases}$$

The mean of Y is then

$$E(Y) = \int_0^{\infty} y3y^{-4}dy = \frac{3}{2}.$$ ◀

Although the method of Example 4.1.11 can be used to find the mean of a continuous random variable, it is not actually necessary to determine the p.d.f. of $r(X)$ in order to calculate the expectation $E[r(X)]$. In fact, it can be shown that the value of $E[r(X)]$ can always be calculated directly using the following result.

Theorem 4.1.1 Law of the Unconscious Statistician. Let X be a random variable, and let r be a real-valued function of a real variable. If X has a continuous distribution, then

$$E[r(X)] = \int_{-\infty}^{\infty} r(x)f(x)\,dx, \tag{4.1.9}$$

if the mean exists. If X has a discrete distribution, then

$$E[r(X)] = \sum_{\text{All } x} r(x)f(x), \tag{4.1.10}$$

if the mean exists.

Proof A general proof will not be given here. However, we shall provide a proof for two special cases. First, suppose that the distribution of X is discrete. Then the distribution of Y must also be discrete. Let g be the p.f. of Y. For this case,

$$\begin{aligned}\sum_y yg(y) &= \sum_y y\,\Pr[r(X) = y] \\ &= \sum_y y \sum_{x:r(x)=y} f(x) \\ &= \sum_y \sum_{x:r(x)=y} r(x)f(x) = \sum_x r(x)f(x).\end{aligned}$$

Hence, Eq. (4.1.10) yields the same value as one would obtain from Definition 4.1.1 applied to Y.

Second, suppose that the distribution of X is continuous. Suppose also, as in Sec. 3.8, that $r(x)$ is either strictly increasing or strictly decreasing with differentiable inverse $s(y)$. Then, if we change variables in Eq. (4.1.9) from x to $y = r(x)$,

$$\int_{-\infty}^{\infty} r(x)f(x)\,dx = \int_{-\infty}^{\infty} yf[s(y)]\left|\frac{ds(y)}{dy}\right|\,dy.$$

It now follows from Eq. (3.8.3) that the right side of this equation is equal to

$$\int_{-\infty}^{\infty} yg(y)\,dy.$$

Hence, Eqs. (4.1.8) and (4.1.9) yield the same value. ■

Theorem 4.1.1 is called the law of the unconscious statistician because many people treat Eqs. (4.1.9) and (4.1.10) as the definition of $E[r(X)]$ and forget that the definition of the mean of $Y = r(X)$ is given in Definitions 4.1.2 and 4.1.4.

Example 4.1.12 Failure Rate and Time to Failure. In Example 4.1.11, we can apply Theorem 4.1.1 to find

$$E(Y) = \int_0^1 \frac{1}{x} 3x^2 dx = \frac{3}{2},$$

the same result we got in Example 4.1.11. ◀

Example 4.1.13 Determining the Expectation of $X^{1/2}$. Suppose that the p.d.f. of X is as given in Example 4.1.6 and that $Y = X^{1/2}$. Then, by Eq. (4.1.9),

$$E(Y) = \int_0^1 x^{1/2}(2x)\, dx = 2\int_0^1 x^{3/2}\, dx = \frac{4}{5}.$$ ◀

Note: In General, $E[g(X)] \neq g(E(X))$**.** In Example 4.1.13, the mean of $X^{1/2}$ is 4/5. The mean of X was computed in Example 4.1.6 as 2/3. Note that $4/5 \neq (2/3)^{1/2}$. In fact, unless g is a linear function, it is generally the case that $E[g(X)] \neq g(E(X))$. A linear function g does satisfy $E[g(X)] = g(E(X))$, as we shall see in Theorem 4.2.1.

Example 4.1.14 Option Pricing. Suppose that common stock in the up-and-coming company A is currently priced at \$200 per share. As an incentive to get you to work for company A, you might be offered an option to buy a certain number of shares of the stock, one year from now, at a price of \$200. This could be quite valuable if you believed that the stock was very likely to rise in price over the next year. For simplicity, suppose that the price X of the stock one year from now is a discrete random variable that can take only two values (in dollars): 260 and 180. Let p be the probability that $X = 260$. You want to calculate the value of these stock options, either because you contemplate the possibility of selling them or because you want to compare Company A's offer to what other companies are offering. Let Y be the value of the option for one share when it expires in one year. Since nobody would pay \$200 for the stock if the price X is less than \$200, the value of the stock option is 0 if $X = 180$. If $X = 260$, one could buy the stock for \$200 per share and then immediately sell it for \$260. This brings in a profit of \$60 per share. (For simplicity, we shall ignore dividends and the transaction costs of buying and selling stocks.) Then $Y = h(X)$ where

$$h(x) = \begin{cases} 0 & \text{if } x = 180, \\ 60 & \text{if } x = 260. \end{cases}$$

Assume that an investor could earn 4% risk-free on any money invested for this same year. (Assume that the 4% includes any compounding.) If no other investment options were available, a fair cost of the option would then be what is called the *present value* of $E(Y)$ in one year. This equals the value c such that $E(Y) = 1.04c$. That is, the expected value of the option equals the amount of money the investor would have after one year without buying the option. We can find $E(Y)$ easily:

$$E(Y) = 0 \times (1 - p) + 60 \times p = 60p.$$

So, the fair price of an option to buy one share would be $c = 60p/1.04 = 57.69p$.

How should one determine the probability p? There is a standard method used in the finance industry for choosing p in this example. That method is to assume that

the present value of the mean of X (the stock price in one year) is equal to the current value of the stock price. That is, assume that the expected value of buying one share of stock and waiting one year to sell is the same as the result of investing the current cost of the stock risk-free for one year (multiplying by 1.04 in this example). In our example, this means $E(X) = 200 \times 1.04$. Since $E(X) = 260p + 180(1 - p)$, we set

$$200 \times 1.04 = 260p + 180(1 - p),$$

and obtain $p = 0.35$. The resulting price of an option to buy one share for \$200 in one year would be $\$57.69 \times 0.35 = \$20.19$. This price is called the *risk-neutral price of the option.* One can prove (see Exercise 14 in this section) that any price other than \$20.19 for the option would lead to unpleasant consequences in the market. ◀

Functions of Several Random Variables

Example 4.1.15 The Expectation of a Function of Two Variables. Let X and Y have a joint p.d.f., and suppose that we want the mean of $X^2 + Y^2$. The most straightforward but most difficult way to do this would be to use the methods of Sec. 3.9 to find the distribution of $Z = X^2 + Y^2$ and then apply the definition of mean to Z. ◀

There is a version of Theorem 4.1.1 for functions of more than one random variable. Its proof is not given here.

Theorem 4.1.2 Law of the Unconscious Statistician. Suppose that $X_1, \dots, X_n$ are random variables with the joint p.d.f. $f(x_1, \dots, x_n)$. Let r be a real-valued function of n real variables, and suppose that $Y = r(X_1, \dots, X_n)$. Then $E(Y)$ can be determined directly from the relation

$$E(Y) = \int \underset{R^n}{\cdots} \int r(x_1, \dots, x_n) f(x_1, \dots, x_n)\, dx_1 \cdots dx_n,$$

if the mean exists. Similarly, if $X_1, \dots, X_n$ have a discrete joint distribution with p.f. $f(x_1, \dots, x_n)$, the mean of $Y = r(X_1, \dots, X_n)$ is

$$E(Y) = \sum_{\text{All } x_1,\dots,x_n} r(x_1, \dots, x_n) f(x_1, \dots, x_n),$$

if the mean exists. ■

Example 4.1.16 Determining the Expectation of a Function of Two Variables. Suppose that a point (X, Y) is chosen at random from the square S containing all points (x, y) such that $0 \le x \le 1$ and $0 \le y \le 1$. We shall determine the expected value of $X^2 + Y^2$.

Since X and Y have the uniform distribution over the square S, and since the area of S is 1, the joint p.d.f. of X and Y is

$$f(x, y) = \begin{cases} 1 & \text{for } (x, y) \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} E(X^2 + Y^2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x^2 + y^2) f(x, y)\, dx\, dy \\ &= \int_0^1 \int_0^1 (x^2 + y^2)\, dx\, dy = \frac{2}{3}. \end{aligned}$$ ◀

Note: More General Distributions. In Example 3.2.7, we introduced a type of distribution that was neither discrete nor continuous. It is possible to define expectations for such distributions also. The definition is rather cumbersome, and we shall not pursue it here.

Summary

The expectation, expected value, or mean of a random variable is a summary of its distribution. If the probability distribution is thought of as a distribution of mass along the real line, then the mean is the center of mass. The mean of a function r of a random variable X can be calculated directly from the distribution of X without first finding the distribution of $r(X)$. Similarly, the mean of a function of a random vector $\boldsymbol{X}$ can be calculated directly from the distribution of $\boldsymbol{X}$.

Exercises

1. Suppose that X has the uniform distribution on the interval $[a, b]$. Find the mean of X.

2. If an integer between 1 and 100 is to be chosen at random, what is the expected value?

3. In a class of 50 students, the number of students n_i of each age i is shown in the following table:

Age i	n_i
18	20
19	22
20	4
21	3
25	1

If a student is to be selected at random from the class, what is the expected value of his age?

4. Suppose that one word is to be selected at random from the sentence THE GIRL PUT ON HER BEAUTIFUL RED HAT. If X denotes the number of letters in the word that is selected, what is the value of $E(X)$?

5. Suppose that one letter is to be selected at random from the 30 letters in the sentence given in Exercise 4. If Y denotes the number of letters in the word in which the selected letter appears, what is the value of $E(Y)$?

6. Suppose that a random variable X has a continuous distribution with the p.d.f. f given in Example 4.1.6. Find the expectation of $1/X$.

7. Suppose that a random variable X has the uniform distribution on the interval [0, 1]. Show that the expectation of $1/X$ is infinite.

8. Suppose that X and Y have a continuous joint distribution for which the joint p.d.f. is as follows:

$$f(x, y) = \begin{cases} 12y^2 & \text{for } 0 \le y \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of $E(XY)$.

9. Suppose that a point is chosen at random on a stick of unit length and that the stick is broken into two pieces at that point. Find the expected value of the length of the longer piece.

10. Suppose that a particle is released at the origin of the xy-plane and travels into the half-plane where $x > 0$. Suppose that the particle travels in a straight line and that the angle between the positive half of the x-axis and this line is α, which can be either positive or negative. Suppose, finally, that the angle α has the uniform distribution on the interval $[-\pi/2, \pi/2]$. Let Y be the ordinate of the point at which the particle hits the vertical line $x = 1$. Show that the distribution of Y is a Cauchy distribution.

11. Suppose that the random variables $X_1, \dots, X_n$ form a random sample of size n from the uniform distribution on the interval [0, 1]. Let $Y_1 = \min\{X_1, \dots, X_n\}$, and let $Y_n = \max\{X_1, \dots, X_n\}$. Find $E(Y_1)$ and $E(Y_n)$.

12. Suppose that the random variables $X_1, \dots, X_n$ form a random sample of size n from a continuous distribution for which the c.d.f. is F, and let the random variables Y_1 and Y_n be defined as in Exercise 11. Find $E[F(Y_1)]$ and $E[F(Y_n)]$.

13. A stock currently sells for \$110 per share. Let the price of the stock at the end of a one-year period be X, which will take one of the values \$100 or \$300. Suppose that you have the option to buy shares of this stock at \$150 per share at the end of that one-year period. Suppose that money

could earn 5.8% risk-free over that one-year period. Find the risk-neutral price for the option to buy one share.

14. Consider the situation of pricing a stock option as in Example 4.1.14. We want to prove that a price other than \$20.19 for the option to buy one share in one year for \$200 would be unfair in some way.

a. Suppose that an investor (who has several shares of the stock already) makes the following transactions. She buys three more shares of the stock at \$200 per share and sells four options for \$20.19 each. The investor must borrow the extra \$519.24 necessary to make these transactions at 4% for the year. At the end of the year, our investor might have to sell four shares for \$200 each to the person who bought the options. In any event, she sells enough stock to pay back the amount borrowed plus the 4 percent interest. Prove that the investor has the same net worth (within rounding error) at the end of the year as she would have had without making these transactions, no matter what happens to the stock price. (A combination of stocks and options that produces no change in net worth is called a *risk-free portfolio*.)

b. Consider the same transactions as in part (a), but this time suppose that the option price is \$$x$ where $x < 20.19$. Prove that our investor loses $|4.16x - 84|$ dollars of net worth no matter what happens to the stock price.

c. Consider the same transactions as in part (a), but this time suppose that the option price is \$$x$ where $x > 20.19$. Prove that our investor gains $4.16x - 84$ dollars of net worth no matter what happens to the stock price.

The situations in parts (b) and (c) are called *arbitrage opportunities*. Such opportunities rarely exist for any length of time in financial markets. Imagine what would happen if the three shares and four options were changed to three million shares and four million options.

15. In Example 4.1.14, we showed how to price an option to buy one share of a stock at a particular price at a particular time in the future. This type of option is called a *call option*. A *put option* is an option to sell a share of a stock at a particular price \$$y$ at a particular time in the future. (If you don't own any shares when you wish to exercise the option, you can always buy one at the market price and then sell it for \$$y$.) The same sort of reasoning as in Example 4.1.14 could be used to price a put option. Consider the same stock as in Example 4.1.14 whose price in one year is X with the same distribution as in the example and the same risk-free interest rate. Find the risk-neutral price for an option to sell one share of that stock in one year at a price of \$220.

16. Let Y be a discrete random variable whose p.f. is the function f in Example 4.1.4. Let $X = |Y|$. Prove that the distribution of X has the p.d.f. in Example 4.1.5.

4.2 Properties of Expectations

In this section, we present some results that simplify the calculation of expectations for some common functions of random variables.

Basic Theorems

Suppose that X is a random variable for which the expectation $E(X)$ exists. We shall present several results pertaining to the basic properties of expectations.

Theorem 4.2.1 Linear Function. If $Y = aX + b$, where a and b are finite constants, then

$$E(Y) = aE(X) + b.$$

Proof We first shall assume, for convenience, that X has a continuous distribution for which the p.d.f. is f. Then

$$\begin{aligned} E(Y) = E(aX + b) &= \int_{-\infty}^{\infty} (ax + b) f(x)\, dx \\ &= a \int_{-\infty}^{\infty} x f(x)\, dx + b \int_{-\infty}^{\infty} f(x)\, dx \\ &= aE(X) + b. \end{aligned}$$

A similar proof can be given for a discrete distribution. ■

Example 4.2.1 **Calculating the Expectation of a Linear Function.** Suppose that $E(X) = 5$. Then

$$E(3X - 5) = 3E(X) - 5 = 10$$

and

$$E(-3X + 15) = -3E(X) + 15 = 0. \qquad \blacktriangleleft$$

The following result follows from Theorem 4.2.1 with $a = 0$.

Corollary 4.2.1 If $X = c$ with probability 1, then $E(X) = c$. ■

Example 4.2.2 **Investment.** An investor is trying to choose between two possible stocks to buy for a three-month investment. One stock costs \$50 per share and has a rate of return of R_1 dollars per share for the three-month period, where R_1 is a random variable. The second stock costs \$30 per share and has a rate of return of R_2 per share for the same three-month period. The investor has a total of \$6000 to invest. For this example, suppose that the investor will buy shares of only one stock. (In Example 4.2.3, we shall consider strategies in which the investor buys more than one stock.) Suppose that R_1 has the uniform distribution on the interval $[-10, 20]$ and that R_2 has the uniform distribution on the interval $[-4.5, 10]$. We shall first compute the expected dollar value of investing in each of the two stocks. For the first stock, the \$6000 will purchase 120 shares, so the return will be $120R_1$, whose mean is $120E(R_1) = 600$. (Solve Exercise 1 in Sec. 4.1 to see why $E(R_1) = 5$.) For the second stock, the \$6000 will purchase 200 shares, so the return will be $200R_2$, whose mean is $200E(R_2) = 550$. The first stock has a higher expected return.

In addition to calculating expected return, we should also ask which of the two investments is riskier. We shall now compute the value at risk (VaR) at probability level 0.97 for each investment. (See Example 3.3.7 on page 113.) VaR will be the negative of the $1 - 0.97 = 0.03$ quantile for the return on each investment. For the first stock, the return $120R_1$ has the uniform distribution on the interval $[-1200, 2400]$ (see Exercise 14 in Sec. 3.8) whose 0.03 quantile is (according to Example 3.3.8 on page 114) $0.03 \times 2400 + 0.97 \times (-1200) = -1092$. So VaR$= 1092$. For the second stock, the return $200R_2$ has the uniform distribution on the interval $[-900, 2000]$ whose 0.03 quantile is $0.03 \times 2000 + 0.97 \times (-900) = -813$. So VaR$= 813$. Even though the first stock has higher expected return, the second stock seems to be slightly less risky in terms of VaR. How should we balance risk and expected return to choose between the two purchases? One way to answer this question is illustrated in Example 4.8.10, after we learn about utility. ◄

Theorem 4.2.2 If there exists a constant such that $\Pr(X \geq a) = 1$, then $E(X) \geq a$. If there exists a constant b such that $\Pr(X \leq b) = 1$, then $E(X) \leq b$.

Proof We shall assume again, for convenience, that X has a continuous distribution for which the p.d.f. is f, and we shall suppose first that $\Pr(X \geq a) = 1$. Because X is bounded below, the second integral in (4.1.5) is finite. Then

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x)\, dx = \int_{a}^{\infty} x f(x)\, dx \\ &\geq \int_{a}^{\infty} a f(x)\, dx = a \Pr(X \geq a) = a. \end{aligned}$$

The proof of the other part of the theorem and the proof for a discrete distribution are similar. ■

It follows from Theorem 4.2.2 that if $\Pr(a \le X \le b) = 1$, then $a \le E(X) \le b$.

Theorem 4.2.3 Suppose that $E(X) = a$ and that either $\Pr(X \ge a) = 1$ or $\Pr(X \le a) = 1$. Then $\Pr(X = a) = 1$.

Proof We shall provide a proof for the case in which X has a discrete distribution and $\Pr(X \ge a) = 1$. The other cases are similar. Let $x_1, x_2, \dots$ include every value $x > a$ such that $\Pr(X = x) > 0$, if any. Let $p_0 = \Pr(X = a)$. Then,

$$E(X) = p_0 a + \sum_{j=1}^{\infty} x_j \Pr(X = x_j). \tag{4.2.1}$$

Each x_j in the sum on the right side of Eq. (4.2.1) is greater than a. If we replace all of the x_j's by a, the sum can't get larger, and hence

$$E(X) \ge p_0 a + \sum_{j=1}^{\infty} a \Pr(X = x_j) = a. \tag{4.2.2}$$

Furthermore, the inequality in Eq. (4.2.2) will be strict if there is even one $x > a$ with $\Pr(X = x) > 0$. This contradicts $E(X) = a$. Hence, there can be no $x > a$ such that $\Pr(X = x) > 0$. ■

Theorem 4.2.4 If $X_1, \dots, X_n$ are n random variables such that each expectation $E(X_i)$ is finite $(i = 1, \dots, n)$, then

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n).$$

Proof We shall first assume that $n = 2$ and also, for convenience, that X_1 and X_2 have a continuous joint distribution for which the joint p.d.f. is f. Then

$$\begin{aligned}
E(X_1 + X_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_1 + x_2) f(x_1, x_2)\, dx_1\, dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f(x_1, x_2)\, dx_1\, dx_2 + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2 f(x_1, x_2)\, dx_1\, dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f(x_1, x_2)\, dx_2\, dx_1 + \int_{-\infty}^{\infty} x_2 f_2(x_2)\, dx_2 \\
&= \int_{-\infty}^{\infty} x_1 f_1(x_1)\, dx_1 + \int_{-\infty}^{\infty} x_2 f_2(x_2)\, dx_2 \\
&= E(X_1) + E(X_2),
\end{aligned}$$

where f_1 and f_2 are the marginal p.d.f.'s of X_1 and X_2. The proof for a discrete distribution is similar. Finally, the theorem can be established for each positive integer n by an induction argument. ■

It should be emphasized that, in accordance with Theorem 4.2.4, the expectation of the sum of several random variables always equals the sum of their individual expectations, regardless of what their joint distribution is. Even though the joint p.d.f. of X_1 and X_2 appeared in the proof of Theorem 4.2.4, only the marginal p.d.f.'s figured into the calculation of $E(X_1 + X_2)$.

The next result follows easily from Theorems 4.2.1 and 4.2.4.

Corollary 4.2.2 Assume that $E(X_i)$ is finite for $i = 1, \dots, n$. For all constants $a_1, \dots, a_n$ and b,

$$E(a_1 X_1 + \cdots + a_n X_n + b) = a_1 E(X_1) + \cdots + a_n E(X_n) + b. \quad ■$$

Example 4.2.3 Investment Portfolio. Suppose that the investor with \$6000 in Example 4.2.2 can buy shares of both of the two stocks. Suppose that the investor buys s_1 shares of the first stock at \$50 per share and s_2 shares of the second stock at \$30 per share. Such a combination of investments is called a *portfolio*. Ignoring possible problems with fractional shares, the values of s_1 and s_2 must satisfy

$$50s_1 + 30s_2 = 6000,$$

in order to invest the entire \$6000. The return on this portfolio will be $s_1R_1 + s_2R_2$. The mean return will be

$$s_1E(R_1) + s_2E(R_2) = 5s_1 + 2.75s_2.$$

For example, if $s_1 = 54$ and $s_2 = 110$, then the mean return is 572.5. ◀

Example 4.2.4 Sampling without Replacement. Suppose that a box contains red balls and blue balls and that the proportion of red balls in the box is p $(0 \le p \le 1)$. Suppose that n balls are selected from the box at random *without replacement*, and let X denote the number of red balls that are selected. We shall determine the value of $E(X)$.

We shall begin by defining n random variables $X_1, \dots, X_n$ as follows: For $i = 1, \dots, n$, let $X_i = 1$ if the ith ball that is selected is red, and let $X_i = 0$ if the ith ball is blue. Since the n balls are selected without replacement, the random variables $X_1, \dots, X_n$ are dependent. However, the marginal distribution of each X_i can be derived easily (see Exercise 10 of Sec. 1.7). We can imagine that all the balls are arranged in the box in some random order, and that the first n balls in this arrangement are selected. Because of randomness, the probability that the ith ball in the arrangement will be red is simply p. Hence, for $i = 1, \dots, n$,

$$\Pr(X_i = 1) = p \quad \text{and} \quad \Pr(X_i = 0) = 1 - p. \tag{4.2.3}$$

Therefore, $E(X_i) = 1(p) + 0(1 - p) = p$.

From the definition of $X_1, \dots, X_n$, it follows that $X_1 + \dots + X_n$ is equal to the total number of red balls that are selected. Therefore, $X = X_1 + \dots + X_n$ and, by Theorem 4.2.4,

$$E(X) = E(X_1) + \dots + E(X_n) = np. \tag{4.2.4}$$

◀

Note: In General, $E[g(X)] \ne g(E(X))$. Theorems 4.2.1 and 4.2.4 imply that if g is a linear function of a random vector $\boldsymbol{X}$, then $E[g(\boldsymbol{X})] = g(E(\boldsymbol{X}))$. For a nonlinear function g, we have already seen Example 4.1.13 in which $E[g(\boldsymbol{X})] \ne g(E(\boldsymbol{X}))$. Jensen's inequality (Theorem 4.2.5) gives a relationship between $E[g(\boldsymbol{X})]$ and $g(E(\boldsymbol{X}))$ for another special class of functions.

Definition 4.2.1 Convex Functions. A function g of a vector argument is *convex* if, for every $\alpha \in (0, 1)$, and every $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$g[\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}] \ge \alpha g(\boldsymbol{x}) + (1-\alpha)g(\boldsymbol{y}).$$

The proof of Theorem 4.2.5 is not given, but one special case is left to the reader in Exercise 13.

Theorem 4.2.5 Jensen's Inequality. Let g be a convex function, and let $\boldsymbol{X}$ be a random vector with finite mean. Then $E[g(\boldsymbol{X})] \ge g(E(\boldsymbol{X}))$. ■

Example 4.2.5 Sampling with Replacement. Suppose again that in a box containing red balls and blue balls, the proportion of red balls is p $(0 \le p \le 1)$. Suppose now, however, that a random sample of n balls is selected from the box *with replacement*. If X denotes the number of red balls in the sample, then X has the binomial distribution with parameters n and p, as described in Sec. 3.1. We shall now determine the value of $E(X)$.

As before, for $i = 1, \ldots, n$, let $X_i = 1$ if the ith ball that is selected is red, and let $X_i = 0$ otherwise. Then, as before, $X = X_1 + \cdots + X_n$. In this problem, the random variables $X_1, \ldots, X_n$ are independent, and the marginal distribution of each X_i is again given by Eq. (4.2.3). Therefore, $E(X_i) = p$ for $i = 1, \ldots, n$, and it follows from Theorem 4.2.4 that

$$E(X) = np. \tag{4.2.5}$$

Thus, the mean of the binomial distribution with parameters n and p is np. The p.f. $f(x)$ of this binomial distribution is given by Eq. (3.1.4), and the mean can be computed directly from the p.f. as follows:

$$E(X) = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x}. \tag{4.2.6}$$

Hence, by Eq. (4.2.5), the value of the sum in Eq. (4.2.6) must be np. ◄

It is seen from Eqs. (4.2.4) and (4.2.5) that the expected number of red balls in a sample of n balls is np, regardless of whether the sample is selected with or without replacement. However, the distribution of the number of red balls is different depending on whether sampling is done with or without replacement (for $n > 1$). For example, $\Pr(X = n)$ is always smaller in Example 4.2.4 where sampling is done without replacement than in Example 4.2.5 where sampling is done with replacement, if $n > 1$. (See Exercise 27 in Sec. 4.9.)

Example 4.2.6 Expected Number of Matches. Suppose that a person types n letters, types the addresses on n envelopes, and then places each letter in an envelope in a random manner. Let X be the number of letters that are placed in the correct envelopes. We shall find the mean of X. (In Sec. 1.10, we did a more difficult calculation with this same example.)

For $i = 1, \ldots, n$, let $X_i = 1$ if the ith letter is placed in the correct envelope, and let $X_i = 0$ otherwise. Then, for $i = 1, \ldots, n$,

$$\Pr(X_i = 1) = \frac{1}{n} \quad \text{and} \quad \Pr(X_i = 0) = 1 - \frac{1}{n}.$$

Therefore,

$$E(X_i) = \frac{1}{n} \quad \text{for } i = 1, \ldots, n.$$

Since $X = X_1 + \cdots + X_n$, it follows that

$$\begin{aligned} E(X) &= E(X_1) + \cdots + E(X_n) \\ &= \frac{1}{n} + \cdots + \frac{1}{n} = 1. \end{aligned}$$

Thus, the expected value of the number of correct matches of letters and envelopes is 1, regardless of the value of n. ◄

Expectation of a Product of Independent Random Variables

Theorem 4.2.6 If $X_1, \dots, X_n$ are n independent random variables such that each expectation $E(X_i)$ is finite ($i = 1, \dots, n$), then

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i).$$

Proof We shall again assume, for convenience, that $X_1, \dots, X_n$ have a continuous joint distribution for which the joint p.d.f. is f. Also, we shall let f_i denote the marginal p.d.f. of X_i ($i = 1, \dots, n$). Then, since the variables $X_1, \dots, X_n$ are independent, it follows that at every point $(x_1, \dots, x_n) \in R^n$,

$$f(x_1, \dots, x_n) = \prod_{i=1}^{n} f_i(x_i).$$

Therefore,

$$\begin{aligned} E\left(\prod_{i=1}^{n} X_i\right) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} x_i\right) f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\prod_{i=1}^{n} x_i f_i(x_i)\right] dx_1 \dots dx_n \\ &= \prod_{i=1}^{n} \int_{-\infty}^{\infty} x_i f_i(x_i)\, dx_i = \prod_{i=1}^{n} E(X_i). \end{aligned}$$

The proof for a discrete distribution is similar. ■

The difference between Theorem 4.2.4 and Theorem 4.2.6 should be emphasized. If it is assumed that each expectation is finite, the expectation of the sum of a group of random variables is *always* equal to the sum of their individual expectations. However, the expectation of the product of a group of random variables is *not* always equal to the product of their individual expectations. If the random variables are *independent*, then this equality will also hold.

Example 4.2.7 **Calculating the Expectation of a Combination of Random Variables.** Suppose that X_1, X_2, and X_3 are independent random variables such that $E(X_i) = 0$ and $E(X_i^2) = 1$ for $i = 1, 2, 3$. We shall determine the value of $E[X_1^2(X_2 - 4X_3)^2]$.

Since X_1, X_2, and X_3 are independent, it follows that the two random variables X_1^2 and $(X_2 - 4X_3)^2$ are also independent. Therefore,

$$\begin{aligned} E[X_1^2(X_2 - 4X_3)^2] &= E(X_1^2)E[(X_2 - 4X_3)^2] \\ &= E(X_2^2 - 8X_2X_3 + 16X_3^2) \\ &= E(X_2^2) - 8E(X_2X_3) + 16E(X_3^2) \\ &= 1 - 8E(X_2)E(X_3) + 16 \\ &= 17. \end{aligned}$$ ◀

Example 4.2.8 **Repeated Filtering.** A filtration process removes a random proportion of particulates in water to which it is applied. Suppose that a sample of water is subjected to this process twice. Let X_1 be the proportion of the particulates that are removed by the first pass. Let X_2 be the proportion of what remains after the first pass that

is removed by the second pass. Assume that X_1 and X_2 are independent random variables with common p.d.f. $f(x) = 4x^3$ for $0 < x < 1$ and $f(x) = 0$ otherwise. Let Y be the proportion of the original particulates that remain in the sample after two passes. Then $Y = (1 - X_1)(1 - X_2)$. Because X_1 and X_2 are independent, so too are $1 - X_1$ and $1 - X_2$. Since $1 - X_1$ and $1 - X_2$ have the same distribution, they have the same mean, call it μ. It follows that Y has mean μ^2. We can find μ as

$$\mu = E(1 - X_1) = \int_0^1 (1 - x_1)4x_1^3 dx_1 = 1 - \frac{4}{5} = 0.2.$$

It follows that $E(Y) = 0.2^2 = 0.04$. ◀

Expectation for Nonnegative Distributions

Theorem 4.2.7 Integer-Valued Random Variables. Let X be a random variable that can take only the values $0, 1, 2, \ldots$. Then

$$E(X) = \sum_{n=1}^{\infty} \Pr(X \geq n). \tag{4.2.7}$$

Proof First, we can write

$$E(X) = \sum_{n=0}^{\infty} n \Pr(X = n) = \sum_{n=1}^{\infty} n \Pr(X = n). \tag{4.2.8}$$

Next, consider the following triangular array of probabilities:

$$\begin{array}{cccc} \Pr(X=1) & \Pr(X=2) & \Pr(X=3) & \cdots \\ & \Pr(X=2) & \Pr(X=3) & \cdots \\ & & \Pr(X=3) & \cdots \\ & & & \ddots \end{array}$$

We can compute the sum of all the elements in this array in two different ways because all of the summands are nonnegative. First, we can add the elements in each column of the array and then add these column totals. Thus, we obtain the value $\sum_{n=1}^{\infty} n \Pr(X = n)$. Second, we can add the elements in each row of the array and then add these row totals. In this way we obtain the value $\sum_{n=1}^{\infty} \Pr(X \geq n)$. Therefore,

$$\sum_{n=1}^{\infty} n \Pr(X = n) = \sum_{n=1}^{\infty} \Pr(X \geq n).$$

Eq. (4.2.7) now follows from Eq. (4.2.8). ■

Example 4.2.9 Expected Number of Trials. Suppose that a person repeatedly tries to perform a certain task until he is successful. Suppose also that the probability of success on each given trial is p $(0 < p < 1)$ and that all trials are independent. If X denotes the number of the trial on which the first success is obtained, then $E(X)$ can be determined as follows.

Since at least one trial is always required, $\Pr(X \geq 1) = 1$. Also, for $n = 2, 3, \ldots$, at least n trials will be required if and only if none of the first $n - 1$ trials results in success. Therefore,

$$\Pr(X \geq n) = (1 - p)^{n-1}.$$

By Eq. (4.2.7), it follows that

$$E(X) = 1 + (1-p) + (1-p)^2 + \cdots = \frac{1}{1-(1-p)} = \frac{1}{p}. \quad ◀$$

Theorem 4.2.7 has a more general version that applies to all nonnegative random variables.

Theorem 4.2.8 General Nonnegative Random Variable. Let X be a nonnegative random variable with c.d.f. F. Then

$$E(X) = \int_0^\infty [1 - F(x)]dx. \tag{4.2.9}$$

■

The proof of Theorem 4.2.8 is left to the reader in Exercises 1 and 2 in Sec. 4.9.

Example 4.2.10 Expected Waiting Time. Let X be the time that a customer spends waiting for service in a queue. Suppose that the c.d.f. of X is

$$F(x) = \begin{cases} 0 & \text{if } x \le 0, \\ 1 - e^{-2x} & \text{if } x > 0. \end{cases}$$

Then the mean of X is

$$E(X) = \int_0^\infty e^{-2x}dx = \frac{1}{2}. \quad ◀$$

Summary

The mean of a linear function of a random vector is the linear function of the mean. In particular, the mean of a sum is the sum of the means. As an example, the mean of the binomial distribution with parameters n and p is np. No such relationship holds in general for nonlinear functions. For independent random variables, the mean of the product is the product of the means.

Exercises

1. Suppose that the return R (in dollars per share) of a stock has the uniform distribution on the interval $[-3, 7]$. Suppose also, that each share of the stock costs \$1.50. Let Y be the net return (total return minus cost) on an investment of 10 shares of the stock. Compute $E(Y)$.

2. Suppose that three random variables X_1, X_2, X_3 form a random sample from a distribution for which the mean is 5. Determine the value of

$$E(2X_1 - 3X_2 + X_3 - 4).$$

3. Suppose that three random variables X_1, X_2, X_3 form a random sample from the uniform distribution on the interval [0, 1]. Determine the value of

$$E[(X_1 - 2X_2 + X_3)^2].$$

4. Suppose that the random variable X has the uniform distribution on the interval [0, 1], that the random variable Y has the uniform distribution on the interval [5, 9], and that X and Y are independent. Suppose also that a rectangle is to be constructed for which the lengths of two adjacent sides are X and Y. Determine the expected value of the area of the rectangle.

5. Suppose that the variables $X_1, \ldots, X_n$ form a random sample of size n from a given continuous distribution on the real line for which the p.d.f. is f. Find the expectation of the number of observations in the sample that fall within a specified interval $a \le x \le b$.

6. Suppose that a particle starts at the origin of the real line and moves along the line in jumps of one unit. For each jump, the probability is p ($0 \le p \le 1$) that the particle will jump one unit to the left and the probability is $1 - p$ that the particle will jump one unit to the right. Find the expected value of the position of the particle after n jumps.

7. Suppose that on each play of a certain game a gambler is equally likely to win or to lose. Suppose that when he wins, his fortune is doubled, and that when he loses, his fortune is cut in half. If he begins playing with a given fortune c, what is the expected value of his fortune after n independent plays of the game?

8. Suppose that a class contains 10 boys and 15 girls, and suppose that eight students are to be selected at random from the class without replacement. Let X denote the number of boys that are selected, and let Y denote the number of girls that are selected. Find $E(X - Y)$.

9. Suppose that the proportion of defective items in a large lot is p, and suppose that a random sample of n items is selected from the lot. Let X denote the number of defective items in the sample, and let Y denote the number of nondefective items. Find $E(X - Y)$.

10. Suppose that a fair coin is tossed repeatedly until a head is obtained for the first time. **(a)** What is the expected number of tosses that will be required? **(b)** What is the expected number of tails that will be obtained before the first head is obtained?

11. Suppose that a fair coin is tossed repeatedly until exactly k heads have been obtained. Determine the expected number of tosses that will be required. *Hint:* Represent the total number of tosses X in the form $X = X_1 + \cdots + X_k$, where X_i is the number of tosses required to obtain the ith head after $i - 1$ heads have been obtained.

12. Suppose that the two return random variables R_1 and R_2 in Examples 4.2.2 and 4.2.3 are independent. Consider the portfolio at the end of Example 4.2.3 with $s_1 = 54$ shares of the first stock and $s_2 = 110$ shares of the second stock.

a. Prove that the change in value X of the portfolio has the p.d.f.

$$f(x) = \begin{cases} 3.87 \times 10^{-7}(x + 1035) & \text{if } -1035 < x < 560, \\ 6.1728 \times 10^{-4} & \text{if } 560 \le x \le 585, \\ 3.87 \times 10^{-7}(2180 - x) & \text{if } 585 < x < 2180, \\ 0 & \text{otherwise.} \end{cases}$$

Hint: Look at Example 3.9.5.

b. Find the value at risk (VaR) at probability level 0.97 for the portfolio.

13. Prove the special case of Theorem 4.2.5 in which the function g is twice continuously differentiable and $\boldsymbol{X}$ is one-dimensional. You may assume that a twice continuously differentiable convex function has nonnegative second derivative. *Hint:* Expand $g(X)$ around its mean using Taylor's theorem with remainder. Taylor's theorem with remainder says that if $g(x)$ has two continuous derivatives g' and g'' at $x = x_0$, then there exists y between x_0 and x such that

$$g(x) = g(x_0) + (x - x_0)g'(x_0) + \frac{(x - x_0)^2}{2} g''(y).$$

4.3 Variance

Although the mean of a distribution is a useful summary, it does not convey very much information about the distribution. For example, a random variable X with mean 2 has the same mean as the constant random variable Y such that $\Pr(Y = 2) = 1$ even if X is not constant. To distinguish the distribution of X from the distribution of Y in this case, it might be useful to give some measure of how spread out the distribution of X is. The variance of X is one such measure. The standard deviation of X is the square root of the variance. The variance also plays an important role in the approximation methods that arise in Chapter 6.

Example 4.3.1 Stock Price Changes. Consider the prices A and B of two stocks at a time one month in the future. Assume that A has the uniform distribution on the interval [25, 35] and B has the uniform distribution on the interval [15, 45]. It is easy to see (from Exercise 1 in Sec. 4.1) that both stocks have a mean price of 30. But the distributions are very different. For example, A will surely be worth at least 25 while $\Pr(B < 25) = 1/3$. But B has more upside potential also. The p.d.f.'s of these two random variables are plotted in Fig. 4.5. ◀

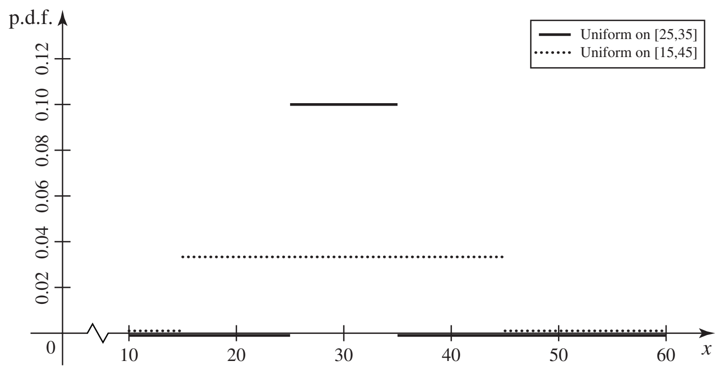


Figure 4.5 The p.d.f.'s of two uniform distributions in Example 4.3.1. Both distributions have mean equal to 30, but they are spread out differently.

Definitions of the Variance and the Standard Deviation

Although the two random prices in Example 4.3.1 have the same mean, price B is more spread out than price A, and it would be good to have a summary of the distribution that makes this easy to see.

Definition 4.3.1 Variance/Standard Deviation. Let X be a random variable with finite mean $\mu = E(X)$. The *variance of* X, denoted by $\text{Var}(X)$, is defined as follows:

$$\text{Var}(X) = E[(X-\mu)^2]. \tag{4.3.1}$$

If X has infinite mean or if the mean of X does not exist, we say that $\text{Var}(X)$ *does not exist*. The *standard deviation of* X is the nonnegative square root of $\text{Var}(X)$ if the variance exists.

If the expectation in Eq. (4.3.1) is infinite, we say that $\text{Var}(X)$ and the standard deviation of X are infinite.

When only one random variable is being discussed, it is common to denote its standard deviation by the symbol σ, and the variance is denoted by σ^2. When more than one random variable is being discussed, the name of the random variable is included as a subscript to the symbol σ, e.g., σ_X would be the standard deviation of X while σ_Y^2 would be the variance of Y.

Example 4.3.2 Stock Price Changes. Return to the two random variables A and B in Example 4.3.1. Using Theorem 4.1.1, we can compute

$$\text{Var}(A) = \int_{25}^{35} (a-30)^2 \frac{1}{10} da = \frac{1}{10}\int_{-5}^{5} x^2 dx = \frac{1}{10}\frac{x^3}{3}\bigg|_{x=-5}^{5} = \frac{25}{3},$$

$$\text{Var}(B) = \int_{15}^{45} (b-30)^2 \frac{1}{30} db = \frac{1}{30}\int_{-15}^{15} y^2 dy = \frac{1}{30}\frac{y^3}{3}\bigg|_{y=-15}^{15} = 75.$$

So, $\text{Var}(B)$ is nine times as large as $\text{Var}(A)$. The standard deviations of A and B are $\sigma_A = 2.87$ and $\sigma_B = 8.66$. ◀

Note: Variance Depends Only on the Distribution. The variance and standard deviation of a random variable X depend only on the distribution of X, just as the expectation of X depends only on the distribution. Indeed, everything that can be computed from the p.f. or p.d.f. depends only on the distribution. Two random

variables with the same distribution will have the same variance, even if they have nothing to do with each other.

Example 4.3.3 Variance and Standard Deviation of a Discrete Distribution. Suppose that a random variable X can take each of the five values −2, 0, 1, 3, and 4 with equal probability. We shall determine the variance and standard deviation of X.

In this example,

$$E(X) = \frac{1}{5}(-2 + 0 + 1 + 3 + 4) = 1.2.$$

Let $\mu = E(X) = 1.2$, and define $W = (X - \mu)^2$. Then $\mathrm{Var}(X) = E(W)$. We can easily compute the p.f. f of W:

x	−2	0	1	3	4
w	10.24	1.44	0.04	3.24	7.84
$f(w)$	1/5	1/5	1/5	1/5	1/5

It follows that

$$\mathrm{Var}(X) = E(W) = \frac{1}{5}[10.24 + 1.44 + 0.04 + 3.24 + 7.84] = 4.56.$$

The standard deviation of X is the square root of the variance, namely, 2.135. ◀

There is an alternative method for calculating the variance of a distribution, which is often easier to use.

Theorem 4.3.1 Alternative Method for Calculating the Variance. For every random variable X, $\mathrm{Var}(X) = E(X^2) - [E(X)]^2$.

Proof Let $E(X) = \mu$. Then

$$\begin{aligned}\mathrm{Var}(X) &= E[(X - \mu)^2]\\ &= E(X^2 - 2\mu X + \mu^2)\\ &= E(X^2) - 2\mu E(X) + \mu^2\\ &= E(X^2) - \mu^2.\end{aligned}$$

■

Example 4.3.4 Variance of a Discrete Distribution. Once again, consider the random variable X in Example 4.3.3, which takes each of the five values −2, 0, 1, 3, and 4 with equal probability. We shall use Theorem 4.3.1 to compute $\mathrm{Var}(X)$. In Example 4.3.3, we computed the mean of X as $\mu = 1.2$. To use Theorem 4.3.1, we need

$$E(X^2) = \frac{1}{5}[(-2)^2 + 0^2 + 1^2 + 3^2 + 4^2] = 6.$$

Because $E(X) = 1.2$, Theorem 4.3.1 says that

$$\mathrm{Var}(X) = 6 - (1.2)^2 = 4.56,$$

which agrees with the calculation in Example 4.3.3. ◀

The variance (as well as the standard deviation) of a distribution provides a measure of the spread or dispersion of the distribution around its mean μ. A small value of the variance indicates that the probability distribution is tightly concentrated around

μ; a large value of the variance typically indicates that the probability distribution has a wide spread around μ. However, the variance of a distribution, as well as its mean, can be made arbitrarily large by placing even a very small but positive amount of probability far enough from the origin on the real line.

Example 4.3.5 Slight Modification of a Bernoulli Distribution. Let X be a discrete random variable with the following p.d.f.:

$$f(x) = \begin{cases} 0.5 & \text{if } x = 0, \\ 0.499 & \text{if } x = 1, \\ 0.001 & \text{if } x = 10{,}000, \\ 0 & \text{otherwise.} \end{cases}$$

There is a sense in which the distribution of X differs very little from the Bernoulli distribution with parameter 0.5. However, the mean and variance of X are quite different from the mean and variance of the Bernoulli distribution with parameter 0.5. Let Y have the Bernoulli distribution with parameter 0.5. In Example 4.1.3, we computed the mean of Y as $E(Y) = 0.5$. Since $Y^2 = Y$, $E(Y^2) = E(Y) = 0.5$, so $\text{Var}(Y) = 0.5 - 0.5^2 = 0.25$. The means of X and X^2 are also straightforward calculations:

$$E(X) = 0.5 \times 0 + 0.499 \times 1 + 0.001 \times 10{,}000 = 10.499$$
$$E(X^2) = 0.5 \times 0^2 + 0.499 \times 1^2 + 0.001 \times 10{,}000^2 = 100{,}000.499.$$

So $\text{Var}(X) = 99{,}890.27$. The mean and variance of X are much larger than the mean and variance of Y. ◀

Properties of the Variance

We shall now present several theorems that state basic properties of the variance. In these theorems we shall assume that the variances of all the random variables exist. The first theorem concerns the possible values of the variance.

Theorem 4.3.2 For each X, $\text{Var}(X) \geq 0$. If X is a bounded random variable, then $\text{Var}(X)$ must exist and be finite.

Proof Because $\text{Var}(X)$ is the mean of a nonnegative random variable $(X - \mu)^2$, it must be nonnegative according to Theorem 4.2.2. If X is bounded, then the mean exists, and hence the variance exists. Furthermore, if X is bounded the so too is $(X - \mu)^2$, so the variance must be finite. ■

The next theorem shows that the variance of a random variable X cannot be 0 unless the entire probability distribution of X is concentrated at a single point.

Theorem 4.3.3 $\text{Var}(X) = 0$ if and only if there exists a constant c such that $\Pr(X = c) = 1$.

Proof Suppose first that there exists a constant c such that $\Pr(X = c) = 1$. Then $E(X) = c$, and $\Pr[(X - c)^2 = 0] = 1$. Therefore,

$$\text{Var}(X) = E[(X - c)^2] = 0.$$

Conversely, suppose that $\text{Var}(X) = 0$. Then $\Pr[(X - \mu)^2 \geq 0] = 1$ but $E[(X - \mu)^2] = 0$. It follows from Theorem 4.2.3 that

$$\Pr[(X - \mu)^2 = 0] = 1.$$

Hence, $\Pr(X = \mu) = 1$. ■

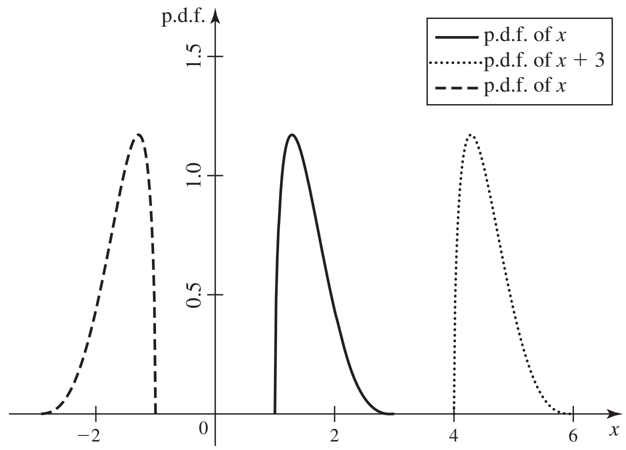


Figure 4.6 The p.d.f. of a random variable X together with the p.d.f.'s of $X + 3$ and $-X$. Note that the spreads of all three distributions appear the same.

Theorem 4.3.4 For constants a and b, let $Y = aX + b$. Then

$$\text{Var}(Y) = a^2 \text{Var}(X),$$

and $\sigma_Y = |a|\sigma_X$.

Proof If $E(X) = \mu$, then $E(Y) = a\mu + b$ by Theorem 4.2.1. Therefore,

$$\begin{aligned}\text{Var}(Y) &= E[(aX + b - a\mu - b)^2] = E[(aX - a\mu)^2] \\ &= a^2 E[(X - \mu)^2] = a^2 \text{Var}(X).\end{aligned}$$

Taking the square root of Var(Y) yields $|a|\sigma_X$. ■

It follows from Theorem 4.3.4 that $\text{Var}(X + b) = \text{Var}(X)$ for every constant b. This result is intuitively plausible, since shifting the entire distribution of X a distance of b units along the real line will change the mean of the distribution by b units but the shift will not affect the dispersion of the distribution around its mean. Figure 4.6 shows the p.d.f. a random variable X together with the p.d.f. of $X + 3$ to illustrate how a shift of the distribution does not affect the spread.

Similarly, it follows from Theorem 4.3.4 that $\text{Var}(-X) = \text{Var}(X)$. This result also is intuitively plausible, since reflecting the entire distribution of X with respect to the origin of the real line will result in a new distribution that is the mirror image of the original one. The mean will be changed from μ to $-\mu$, but the total dispersion of the distribution around its mean will not be affected. Figure 4.6 shows the p.d.f. of a random variable X together with the p.d.f. of $-X$ to illustrate how a reflection of the distribution does not affect the spread.

Example 4.3.6 **Calculating the Variance and Standard Deviation of a Linear Function.** Consider the same random variable X as in Example 4.3.3, which takes each of the five values −2, 0, 1, 3, and 4 with equal probability. We shall determine the variance and standard deviation of $Y = 4X - 7$.

In Example 4.3.3, we computed the mean of X as $\mu = 1.2$ and the variance as 4.56. By Theorem 4.3.4,

$$\text{Var}(Y) = 16\,\text{Var}(X) = 72.96.$$

Also, the standard deviation σ of Y is

$$\sigma_Y = 4\sigma_X = 4(4.56)^{1/2} = 8.54.$$ ◀

The next theorem provides an alternative method for calculating the variance of a sum of independent random variables.

Theorem 4.3.5 If $X_1, \dots, X_n$ are independent random variables with finite means, then

$$\text{Var}(X_1 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n).$$

Proof Suppose first that $n = 2$. If $E(X_1) = \mu_1$ and $E(X_2) = \mu_2$, then

$$E(X_1 + X_2) = \mu_1 + \mu_2.$$

Therefore,

$$\begin{aligned} \text{Var}(X_1 + X_2) &= E[(X_1 + X_2 - \mu_1 - \mu_2)^2] \\ &= E[(X_1 - \mu_1)^2 + (X_2 - \mu_2)^2 + 2(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= \text{Var}(X_1) + \text{Var}(X_2) + 2E[(X_1 - \mu_1)(X_2 - \mu_2)]. \end{aligned}$$

Since X_1 and X_2 are independent,

$$\begin{aligned} E[(X_1 - \mu_1)(X_2 - \mu_2)] &= E(X_1 - \mu_1)E(X_2 - \mu_2) \\ &= (\mu_1 - \mu_1)(\mu_2 - \mu_2) \\ &= 0. \end{aligned}$$

It follows, therefore, that

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2).$$

The theorem can now be established for each positive integer n by an induction argument. ■

It should be emphasized that the random variables in Theorem 4.3.5 must be independent. The variance of the sum of random variables that are not independent will be discussed in Sec. 4.6. By combining Theorems 4.3.4 and 4.3.5, we can now obtain the following corollary.

Corollary 4.3.1 If $X_1, \dots, X_n$ are independent random variables with finite means, and if $a_1, \dots, a_n$ and b are arbitrary constants, then

$$\text{Var}(a_1X_1 + \cdots + a_nX_n + b) = a_1^2\,\text{Var}(X_1) + \cdots + a_n^2\,\text{Var}(X_n). \quad ■$$

Example 4.3.7 Investment Portfolio. An investor with \$100,000 to invest wishes to construct a portfolio consisting of shares of one or both of two available stocks and possibly some fixed-rate investments. Suppose that the two stocks have random rates of return R_1 and R_2 per share for a period of one year. Suppose that R_1 has a distribution with mean 6 and variance 55, while R_2 has mean 4 and variance 28. Suppose that the first stock costs \$60 per share and the second costs \$48 per share. Suppose that money can also be invested at a fixed rate of 3.6 percent per year. The portfolio will consist of s_1 shares of the first stock, s_2 shares of the second stock, and all remaining money (\$$s_3$) invested at the fixed rate. The return on this portfolio will be

$$s_1R_1 + s_2R_2 + 0.036s_3,$$

where the coefficients are constrained by

$$60s_1 + 48s_2 + s_3 = 100{,}000, \tag{4.3.2}$$

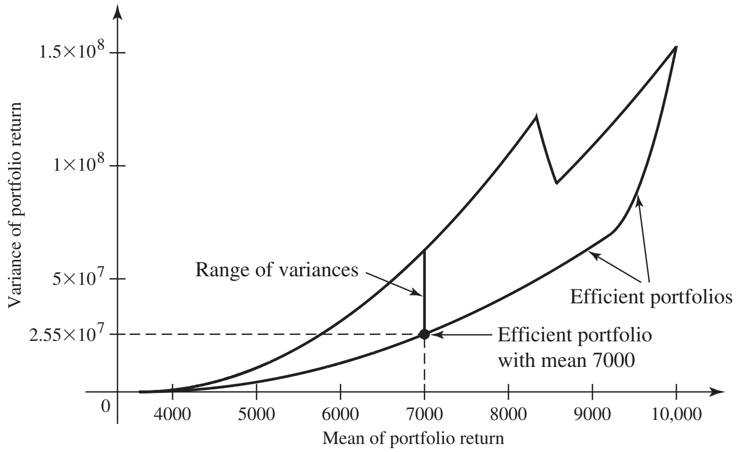


Figure 4.7 The set of all means and variances of investment portfolios in Example 4.3.7. The solid vertical line shows the range of possible variances for portfoloios with a mean of 7000.

as well as $s_1, s_2, s_3 \geq 0$. For now, we shall assume that R_1 and R_2 are independent. The mean and the variance of the return on the portfolio will be

$$E(s_1R_1 + s_2R_2 + 0.036s_3) = 6s_1 + 4s_2 + 0.036s_3,$$
$$\mathrm{Var}(s_1R_1 + s_2R_2 + 0.036s_3) = 55s_1^2 + 28s_2^2.$$

One method for comparing a class of portfolios is to say that portfolio A is at least as good as portfolio B if the mean return for A is at least as large as the mean return for B and if the variance for A is no larger than the variance of B. (See Markowitz, 1987, for a classic treatment of such methods.) The reason for preferring smaller variance is that large variance is associated with large deviations from the mean, and for portfolios with a common mean, some of the large deviations are going to have to be below the mean, leading to the risk of large losses. Figure 4.7 is a plot of the pairs (mean, variance) for all of the possible portfolios in this example. That is, for each (s_1, s_2, s_3) that satisfy (4.3.2), there is a point in the outlined region of Fig. 4.7. The points to the right and toward the bottom are those that have the largest mean return for a fixed variance, and the ones that have the smallest variance for a fixed mean return. These portfolios are called *efficient*. For example, suppose that the investor would like a mean return of 7000. The vertical line segment above 7000 on the horizontal axis in Fig. 4.7 indicates the possible variances of all portfolios with mean return of 7000. Among these, the portfolio with the smallest variance is efficient and is indicated in Fig. 4.7. This portfolio has $s_1 = 524.7$, $s_2 = 609.7$, $s_3 = 39{,}250$, and variance 2.55×10^7. So, every portfolio with mean return greater than 7000 must have variance larger than 2.55×10^7, and every portfolio with variance less than 2.55×10^7 must have mean return smaller than 7000. ◀

The Variance of a Binomial Distribution

We shall now consider again the method of generating a binomial distribution presented in Sec. 4.2. Suppose that a box contains red balls and blue balls, and that the proportion of red balls is p $(0 \leq p \leq 1)$. Suppose also that a random sample of n balls is selected from the box with replacement. For $i = 1, \dots, n$, let $X_i = 1$ if the ith ball that is selected is red, and let $X_i = 0$ otherwise. If X denotes the total number of red balls in the sample, then $X = X_1 + \cdots + X_n$ and X will have the binomial distribution with parameters n and p.

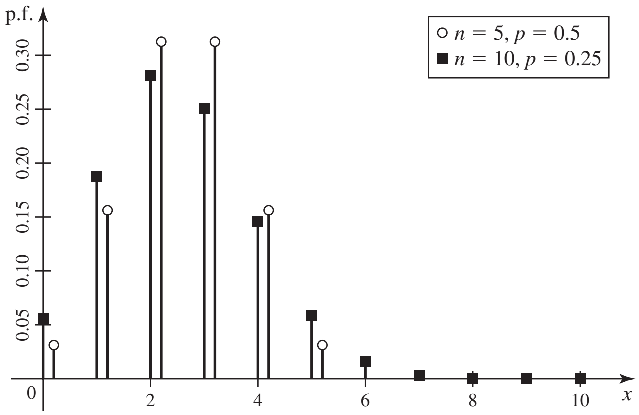


Figure 4.8 Two binomial distributions with the same mean (2.5) but different variances.

Since $X_1, \dots, X_n$ are independent, it follows from Theorem 4.3.5 that

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

According to Example 4.1.3, $E(X_i) = p$ for $i = 1, \dots, n$. Since $X_i^2 = X_i$ for each i, $E(X_i^2) = E(X_i) = p$. Therefore, by Theorem 4.3.1,

$$\begin{aligned}\mathrm{Var}(X_i) &= E(X_i^2) - [E(X_i)]^2 \\ &= p - p^2 = p(1-p).\end{aligned}$$

It now follows that

$$\mathrm{Var}(X) = np(1-p). \tag{4.3.3}$$

Figure 4.8 compares two different binomial distributions with the same mean (2.5) but different variances (1.25 and 1.875). One can see how the p.f. of the distribution with the larger variance ($n = 10$, $p = 0.25$) is higher at more extreme values and lower at more central values than is the p.f. of the distribution with the smaller variance ($n = 5$, $p = 0.5$). Similarly, Fig. 4.5 compares two uniform distributions with the same mean (30) and different variances (8.33 and 75). The same pattern appears, namely that the distribution with larger variance has higher p.d.f. at more extreme values and lower p.d.f. at more central values.

Interquartile Range

Example 4.3.8 **The Cauchy Distribution.** In Example 4.1.8, we saw a distribution (the Cauchy distribution) whose mean did not exist, and hence its variance does not exist. But, we might still want to describe how spread out such a distribution is. For example, if X has the Cauchy distribution and $Y = 2X$, it stands to reason that Y is twice as spread out as X is, but how do we quantify this? ◀

There is a measure of spread that exists for every distribution, regardless of whether or not the distribution has a mean or variance. Recall from Definition 3.3.2 that the quantile function for a random variable is the inverse of the c.d.f., and it is defined for every random variable.

Definition 4.3.2 Interquartile Range (IQR). Let X be a random variable with quantile function $F^{-1}(p)$ for $0 < p < 1$. The *interquartile range (IQR)* is defined to be $F^{-1}(0.75) - F^{-1}(0.25)$.

In words, the IQR is the length of the interval that contains the middle half of the distribution.

Example 4.3.9 The Cauchy Distribution. Let X have the Cauchy distribution. The c.d.f. F of X can be found using a trigonometric substitution in the following integral:

$$F(x) = \int_{-\infty}^{x} \frac{dy}{\pi(1+y^2)} = \frac{1}{2} + \frac{\tan^{-1}(x)}{\pi},$$

where $\tan^{-1}(x)$ is the principal inverse of the tangent function, taking values from $-\pi/2$ to $\pi/2$ as x runs from $-\infty$ to ∞. The quantile function of X is then $F^{-1}(p) = \tan[\pi(p - 1/2)]$ for $0 < p < 1$. The IQR is

$$F^{-1}(0.75) - F^{-1}(0.25) = \tan(\pi/4) - \tan(-\pi/4) = 2.$$

It is not difficult to show that, if $Y = 2X$, then the IQR of Y is 4. (See Exercise 14.) ◀

Summary

The variance of X, denoted by $\mathrm{Var}(X)$, is the mean of $[X - E(X)]^2$ and measures how spread out the distribution of X is. The variance also equals $E(X^2) - [E(X)]^2$. The standard deviation is the square root of the variance. The variance of $aX + b$, where a and b are constants, is $a^2\,\mathrm{Var}(X)$. The variance of the sum of independent random variables is the sum of the variances. As an example, the variance of the binomial distribution with parameters n and p is $np(1 - p)$. The interquartile range (IQR) is the difference between the 0.75 and 0.25 quantiles. The IQR is a measure of spread that exists for every distribution.

Exercises

1. Suppose that X has the uniform distribution on the interval [0, 1]. Compute the variance of X.

2. Suppose that one word is selected at random from the sentence THE GIRL PUT ON HER BEAUTIFUL RED HAT. If X denotes the number of letters in the word that is selected, what is the value of $\mathrm{Var}(X)$?

3. For all numbers a and b such that $a < b$, find the variance of the uniform distribution on the interval $[a, b]$.

4. Suppose that X is a random variable for which $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Show that $E[X(X - 1)] = \mu(\mu - 1) + \sigma^2$.

5. Let X be a random variable for which $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, and let c be an arbitrary constant. Show that

$$E[(X - c)^2] = (\mu - c)^2 + \sigma^2.$$

6. Suppose that X and Y are independent random variables whose variances exist and such that $E(X) = E(Y)$. Show that

$$E[(X - Y)^2] = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

7. Suppose that X and Y are independent random variables for which $\mathrm{Var}(X) = \mathrm{Var}(Y) = 3$. Find the values of **(a)** $\mathrm{Var}(X - Y)$ and **(b)** $\mathrm{Var}(2X - 3Y + 1)$.

8. Construct an example of a distribution for which the mean is finite but the variance is infinite.

9. Let X have the discrete uniform distribution on the integers $1, \dots, n$. Compute the variance of X. *Hint:* You may wish to use the formula $\sum_{k=1}^{n} k^2 = n(n+1)\cdot(2n+1)/6$.

10. Consider the example efficient portfolio at the end of Example 4.3.7. Suppose that R_i has the uniform distribution on the interval $[a_i, b_i]$ for $i = 1, 2$.

a. Find the two intervals $[a_1, b_1]$ and $[a_2, b_2]$. *Hint:* The intervals are determined by the means and variances.

b. Find the value at risk (VaR) for the example portfolio at probability level 0.97. *Hint:* Review Example 3.9.5 to see how to find the p.d.f. of the sum of two uniform random variables.

11. Let X have the uniform distribution on the interval [0, 1]. Find the IQR of X.

12. Let X have the p.d.f. $f(x) = \exp(-x)$ for $x \ge 0$, and $f(x) = 0$ for $x < 0$. Find the IQR of X.

13. Let X have the binomial distribution with parameters 5 and 0.3. Find the IQR of X. *Hint:* Return to Example 3.3.9 and Table 3.1.

14. Let X be a random variable whose interquartile range is η. Let $Y = 2X$. Prove that the interquartile range of Y is 2η.

4.4 Moments

For a random variable X, the means of powers X^k (called moments) for $k > 2$ have useful theoretical properties, and some of them are used for additional summaries of a distribution. The moment generating function is a related tool that aids in deriving distributions of sums of independent random variables and limiting properties of distributions.

Existence of Moments

For each random variable X and every positive integer k, the expectation $E(X^k)$ is called the kth *moment of X*. In particular, in accordance with this terminology, the mean of X is the first moment of X.

It is said that the kth moment exists if and only if $E(|X|^k) < \infty$. If the random variable X is bounded, that is, if there are finite numbers a and b such that $\Pr(a \le X \le b) = 1$, then all moments of X must necessarily exist. It is possible, however, that all moments of X exist even though X is not bounded. It is shown in the next theorem that if the kth moment of X exists, then all moments of lower order must also exist.

Theorem 4.4.1 If $E(|X|^k) < \infty$ for some positive integer k, then $E(|X|^j) < \infty$ for every positive integer j such that $j < k$.

Proof We shall assume, for convenience, that the distribution of X is continuous and the p.d.f. is f. Then

$$\begin{aligned} E(|X|^j) &= \int_{-\infty}^{\infty} |x|^j f(x)\, dx \\ &= \int_{|x|\le 1} |x|^j f(x)\, dx + \int_{|x|>1} |x|^j f(x)\, dx \\ &\le \int_{|x|\le 1} 1 \cdot f(x)\, dx + \int_{|x|>1} |x|^k f(x)\, dx \\ &\le \Pr(|X| \le 1) + E(|X|^k). \end{aligned}$$

By hypothesis, $E(|X|^k) < \infty$. It therefore follows that $E(|X|^j) < \infty$. A similar proof holds for a discrete or a more general type of distribution. ∎

In particular, it follows from Theorem 4.4.1 that if $E(X^2) < \infty$, then both the mean of X and the variance of X exist. Theorem 4.4.1 extends to the case in which

j and k are arbitrary positive numbers rather than just integers. (See Exercise 15 in this section.) We will not make use of such a result in this text, however.

Central Moments Suppose that X is a random variable for which $E(X) = \mu$. For every positive integer k, the expectation $E[(X - \mu)^k]$ is called the kth *central moment of X* or the kth *moment of X about the mean*. In particular, in accordance with this terminology, the variance of X is the second central moment of X.

For every distribution, the first central moment must be 0 because

$$E(X - \mu) = \mu - \mu = 0.$$

Furthermore, if the distribution of X is symmetric with respect to its mean μ, and if the central moment $E[(X - \mu)^k]$ exists for a given odd integer k, then the value of $E[(X - \mu)^k]$ will be 0 because the positive and negative terms in this expectation will cancel one another.

Example 4.4.1 A Symmetric p.d.f. Suppose that X has a continuous distribution for which the p.d.f. has the following form:

$$f(x) = ce^{-(x-3)^2/2} \quad \text{for } -\infty < x < \infty.$$

We shall determine the mean of X and all the central moments.

It can be shown that for every positive integer k,

$$\int_{-\infty}^{\infty} |x|^k e^{-(x-3)^2/2}\, dx < \infty.$$

Hence, all the moments of X exist. Furthermore, since $f(x)$ is symmetric with respect to the point $x = 3$, then $E(X) = 3$. Because of this symmetry, it also follows that $E[(X - 3)^k] = 0$ for every odd positive integer k. For even $k = 2n$, we can find a recursive formula for the sequence of central moments. First, let $y = x - \mu$ in all the integral fomulas. Then, for $n \geq 1$, the $2n$th central moment is

$$m_{2n} = \int_{-\infty}^{\infty} y^{2n} ce^{-y^2/2} dy.$$

Use integration by parts with $u = y^{2n-1}$ and $dv = ye^{-y^2/2}dy$. It follows that $du = (2n - 1)y^{2n-2}dy$ and $v = -e^{-y^2/2}$. So,

$$\begin{aligned} m_{2n} &= \int_{-\infty}^{\infty} u\,dv = uv\Big|_{y=-\infty}^{\infty} - \int_{-\infty}^{\infty} v\,du \\ &= -y^{2n-1}e^{-y^2/2}\Big|_{y=-\infty}^{\infty} + (2n-1)\int_{-\infty}^{\infty} y^{2n-2} ce^{-y^2/2} dy \\ &= (2n-1)m_{2(n-1)}. \end{aligned}$$

Because $y^0 = 1$, m_0 is just the integral of the p.d.f.; hence, $m_0 = 1$. It follows that $m_{2n} = \prod_{i=1}^{n}(2i - 1)$ for $n = 1, 2, \ldots$. So, for example, $m_2 = 1$, $m_4 = 3$, $m_6 = 15$, and so on. ◀

Skewness In Example 4.4.1, we saw that the odd central moments are all 0 for a distribution that is symmetric. This leads to the following distributional summary that is used to measure lack of symmetry.

Definition 4.4.1 Skewness. Let X be a random variable with mean μ, standard deviation σ, and finite third moment. The *skewness* of X is defined to be $E[(X - \mu)^3]/\sigma^3$.

The reason for dividing the third central moment by σ^3 is to make the skewness measure only the lack of symmetry rather than the spread of the distribution.

Example 4.4.2

Skewness of Binomial Distributions. Let X have the binomial distribution with parameters 10 and 0.25. The p.f. of this distribution appears in Fig. 4.8. It is not difficult to see that the p.f. is not symmetric. The skewness can be computed as follows: First, note that the mean is $\mu = 10 \times 0.25 = 2.5$ and that the standard deviation is

$$\sigma = (10 \times 0.25 \times 0.75)^{1/2} = 1.369.$$

Second, compute

$$E[(X - 2.5)^3] = (0 - 2.5)^3 \binom{10}{0} 0.25^0\, 0.75^{10} + \cdots + (10 - 2.5)^3 \binom{10}{10} 0.25^{00}\, 0.75^0$$
$$= 0.9375.$$

Finally, the skewness is

$$\frac{0.9375}{1.369^3} = 0.3652.$$

For comparison, the skewness of the binomial distribution with parameters 10 and 0.2 is 0.4743, and the skewness of the binomial distribution with parameters 10 and 0.3 is 0.2761. The absolute value of the skewness increases as the probability of success moves away from 0.5. It is straightforward to show that the skewness of the binomial distribution with parameters n and p is the negative of the skewness of the binomial distribution with parameters n and $1 - p$. (See Exercise 16 in this section.) ◀

Moment Generating Functions

We shall now consider a different way to characterize the distribution of a random variable that is more closely related to its moments than to where its probability is distributed.

Definition 4.4.2

Moment Generating Function. Let X be a random variable. For each real number t, define

$$\psi(t) = E(e^{tX}). \tag{4.4.1}$$

The function $\psi(t)$ is called the *moment generating function* (abbreviated m.g.f.) of X.

Note: The Moment Generating Function of X Depends Only on the Distribution of X. Since the m.g.f. is the expected value of a function of X, it must depend only on the distribution of X. If X and Y have the same distribution, they must have the same m.g.f.

If the random variable X is bounded, then the expectation in Eq. (4.4.1) must be finite for all values of t. In this case, therefore, the m.g.f. of X will be finite for all values of t. On the other hand, if X is not bounded, then the m.g.f. might be finite for some values of t and might not be finite for others. It can be seen from Eq. (4.4.1), however, that for every random variable X, the m.g.f. $\psi(t)$ must be finite at the point $t = 0$ and at that point its value must be $\psi(0) = E(1) = 1$.

The next result explains how the name "moment generating function" arose.

Theorem 4.4.2

Let X be a random variables whose m.g.f. $\psi(t)$ is finite for all values of t in some open interval around the point $t = 0$. Then, for each integer $n > 0$, the nth moment of X,

$E(X^n)$, is finite and equals the nth derivative $\psi^{(n)}(t)$ at $t = 0$. That is, $E(X^n) = \psi^{(n)}(0)$ for $n = 1, 2, \ldots$.

We sketch the proof at the end of this section.

Example 4.4.3 Calculating an m.g.f. Suppose that X is a random variable for which the p.d.f. is as follows:

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

We shall determine the m.g.f. of X and also $\text{Var}(X)$.

For each real number t,

$$\psi(t) = E(e^{tX}) = \int_0^\infty e^{tx}e^{-x}\,dx$$
$$= \int_0^\infty e^{(t-1)x}\,dx.$$

The final integral in this equation will be finite if and only if $t < 1$. Therefore, $\psi(t)$ is finite only for $t < 1$. For each such value of t,

$$\psi(t) = \frac{1}{1-t}.$$

Since $\psi(t)$ is finite for all values of t in an open interval around the point $t = 0$, all moments of X exist. The first two derivatives of ψ are

$$\psi'(t) = \frac{1}{(1-t)^2} \quad \text{and} \quad \psi''(t) = \frac{2}{(1-t)^3}.$$

Therefore, $E(X) = \psi'(0) = 1$ and $E(X^2) = \psi''(0) = 2$. It now follows that

$$\text{Var}(X) = \psi''(0) - [\psi'(0)]^2 = 1.$$ ◀

Properties of Moment Generating Functions

We shall now present three basic theorems pertaining to moment generating functions.

Theorem 4.4.3 Let X be a random variable for which the m.g.f. is ψ_1; let $Y = aX + b$, where a and b are given constants; and let ψ_2 denote the m.g.f. of Y. Then for every value of t such that $\psi_1(at)$ is finite,

$$\psi_2(t) = e^{bt}\psi_1(at). \tag{4.4.2}$$

Proof By the definition of an m.g.f.,

$$\psi_2(t) = E(e^{tY}) = E[e^{t(aX+b)}] = e^{bt}E(e^{atX}) = e^{bt}\psi_1(at).$$ ■

Example 4.4.4 Calculating the m.g.f. of a Linear Function. Suppose that the distribution of X is as specified in Example 4.4.3. We saw that the m.g.f. of X for $t < 1$ is

$$\psi_1(t) = \frac{1}{1-t}.$$

If $Y = 3 - 2X$, then the m.g.f. of Y is finite for $t > -1/2$ and will have the value

$$\psi_2(t) = e^{3t}\psi_1(-2t) = \frac{e^{3t}}{1+2t}.$$ ◀

The next theorem shows that the m.g.f. of the sum of an arbitrary number of independent random variables has a very simple form. Because of this property, the m.g.f. is an important tool in the study of such sums.

Theorem 4.4.4 Suppose that $X_1, \dots, X_n$ are n independent random variables; and for $i = 1, \dots, n$, let ψ_i denote the m.g.f. of X_i. Let $Y = X_1 + \cdots + X_n$, and let the m.g.f. of Y be denoted by ψ. Then for every value of t such that $\psi_i(t)$ is finite for $i = 1, \dots, n$,

$$\psi(t) = \prod_{i=1}^{n} \psi_i(t). \tag{4.4.3}$$

Proof By definition,

$$\psi(t) = E(e^{tY}) = E[e^{t(X_1+\cdots+X_n)}] = E\left(\prod_{i=1}^{n} e^{tX_i}\right).$$

Since the random variables $X_1, \dots, X_n$ are independent, it follows from Theorem 4.2.6 that

$$E\left(\prod_{i=1}^{n} e^{tX_i}\right) = \prod_{i=1}^{n} E(e^{tX_i}).$$

Hence,

$$\psi(t) = \prod_{i=1}^{n} \psi_i(t). \qquad \blacksquare$$

The Moment Generating Function for the Binomial Distribution Suppose that a random variable X has the binomial distribution with parameters n and p. In Sections 4.2 and 4.3, the mean and the variance of X were determined by representing X as the sum of n independent random variables $X_1, \dots, X_n$. In this representation, the distribution of each variable X_i is as follows:

$$\Pr(X_i = 1) = p \quad \text{and} \quad \Pr(X_i = 0) = 1 - p.$$

We shall now use this representation to determine the m.g.f. of $X = X_1 + \cdots + X_n$.

Since each of the random variables $X_1, \dots, X_n$ has the same distribution, the m.g.f. of each variable will be the same. For $i = 1, \dots, n$, the m.g.f. of X_i is

$$\begin{aligned}\psi_i(t) &= E(e^{tX_i}) = (e^t)\Pr(X_i = 1) + (1)\Pr(X_i = 0)\\ &= pe^t + 1 - p.\end{aligned}$$

It follows from Theorem 4.4.4 that the m.g.f. of X in this case is

$$\psi(t) = (pe^t + 1 - p)^n. \tag{4.4.4}$$

Uniqueness of Moment Generating Functions We shall now state one more important property of the m.g.f. The proof of this property is beyond the scope of this book and is omitted.

Theorem 4.4.5 If the m.g.f.'s of two random variables X_1 and X_2 are finite and identical for all values of t in an open interval around the point $t = 0$, then the probability distributions of X_1 and X_2 must be identical. ■

Theorem 4.4.5 is the justification for the claim made at the start of this discussion, namely, that the m.g.f. is another way to characterize the distribution of a random variable.

The Additive Property of the Binomial Distribution Moment generating functions provide a simple way to derive the distribution of the sum of two independent binomial random variables with the same second parameter.

Theorem 4.4.6 If X_1 and X_2 are independent random variables, and if X_i has the binomial distribution with parameters n_i and p $(i = 1, 2)$, then $X_1 + X_2$ has the binomial distribution with parameters $n_1 + n_2$ and p.

Proof L et ψ_i denote the m.g.f. of X_i for $i = 1, 2$. It follows from Eq. (4.4.4) that

$$\psi_i(t) = (pe^t + 1 - p)^{n_i}.$$

Let ψ denote the m.g.f. of $X_1 + X_2$. Then, by Theorem 4.4.4,

$$\psi(t) = (pe^t + 1 - p)^{n_1+n_2}.$$

It can be seen from Eq. (4.4.4) that this function ψ is the m.g.f. of the binomial distribution with parameters $n_1 + n_2$ and p. Hence, by Theorem 4.4.5, the distribution of $X_1 + X_2$ must be that binomial distribution. ∎

Sketch of the Proof of Theorem 4.4.2

First, we indicate why all moments of X are finite. Let $t > 0$ be such that both $\psi(t)$ and $\psi(-t)$ are finite. Define $g(x) = e^{tx} + e^{-tx}$. Notice that

$$E[g(X)] = \psi(t) + \psi(-t) < \infty. \tag{4.4.5}$$

On every bounded interval of x values, $g(x)$ is bounded. For each integer $n > 0$, as $|x| \to \infty$, $g(x)$ is eventually larger than $|x|^n$. It follows from these facts and (4.4.5) that $E|X^n| < \infty$.

Although it is beyond the scope of this book, it can be shown that the derivative $\psi'(t)$ exists at the point $t = 0$, and that at $t = 0$, the derivative of the expectation in Eq. (4.4.1) must be equal to the expectation of the derivative. Thus,

$$\psi'(0) = \left[\frac{d}{dt}E(e^{tX})\right]_{t=0} = E\left[\left(\frac{d}{dt}e^{tX}\right)_{t=0}\right].$$

But

$$\left(\frac{d}{dt}e^{tX}\right)_{t=0} = (Xe^{tX})_{t=0} = X.$$

It follows that

$$\psi'(0) = E(X).$$

In other words, the derivative of the m.g.f. $\psi(t)$ at $t = 0$ is the mean of X.

Furthermore, it can be shown that it is possible to differentiate $\psi(t)$ an arbitrary number of times at the point $t = 0$. For $n = 1, 2, \ldots,$ the nth derivative $\psi^{(n)}(0)$ at $t = 0$ will satisfy the following relation:

$$\psi^{(n)}(0) = \left[\frac{d^n}{dt^n}E(e^{tX})\right]_{t=0} = E\left[\left(\frac{d^n}{dt^n}e^{tX}\right)_{t=0}\right]$$
$$= E[(X^n e^{tX})_{t=0}] = E(X^n).$$

Thus, $\psi'(0) = E(X)$, $\psi''(0) = E(X^2)$, $\psi'''(0) = E(X^3)$, and so on. Hence, we see that the m.g.f., if it is finite in an open interval around $t = 0$, can be used to generate all of the moments of the distribution by taking derivatives at $t = 0$.

Summary

If the kth moment of a random variable exists, then so does the jth moment for every $j < k$. The moment generating function of X, $\psi(t) = E(e^{tX})$, if it is finite for t in a neighborhood of 0, can be used to find moments of X. The kth derivative of $\psi(t)$ at $t = 0$ is $E(X^k)$. The m.g.f. characterizes the distribution in the sense that all random variables that have the same m.g.f. have the same distribution.

Exercises

1. If X has the uniform distribution on the interval $[a, b]$, what is the value of the fifth central moment of X?

2. If X has the uniform distribution on the interval $[a, b]$, write a formula for every even central moment of X.

3. Suppose that X is a random variable for which $E(X) = 1$, $E(X^2) = 2$, and $E(X^3) = 5$. Find the value of the third central moment of X.

4. Suppose that X is a random variable such that $E(X^2)$ is finite. **(a)** Show that $E(X^2) \ge [E(X)]^2$. **(b)** Show that $E(X^2) = [E(X)]^2$ if and only if there exists a constant c such that $\Pr(X = c) = 1$. *Hint:* $\text{Var}(X) \ge 0$.

5. Suppose that X is a random variable with mean μ and variance σ^2, and that the fourth moment of X is finite. Show that

$$E[(X - \mu)^4] \ge \sigma^4.$$

6. Suppose that X has the uniform distribution on the interval $[a, b]$. Determine the m.g.f. of X.

7. Suppose that X is a random variable for which the m.g.f. is as follows:

$$\psi(t) = \frac{1}{4}(3e^t + e^{-t}) \quad \text{for } -\infty < t < \infty.$$

Find the mean and the variance of X.

8. Suppose that X is a random variable for which the m.g.f. is as follows:

$$\psi(t) = e^{t^2+3t} \quad \text{for } -\infty < t < \infty.$$

Find the mean and the variance of X.

9. Let X be a random variable with mean μ and variance σ^2, and let $\psi_1(t)$ denote the m.g.f. of X for $-\infty < t < \infty$. Let c be a given positive constant, and let Y be a random variable for which the m.g.f. is

$$\psi_2(t) = e^{c[\psi_1(t)-1]} \quad \text{for } -\infty < t < \infty.$$

Find expressions for the mean and the variance of Y in terms of the mean and the variance of X.

10. Suppose that the random variables X and Y are i.i.d. and that the m.g.f. of each is

$$\psi(t) = e^{t^2+3t} \quad \text{for } -\infty < t < \infty.$$

Find the m.g.f. of $Z = 2X - 3Y + 4$.

11. Suppose that X is a random variable for which the m.g.f. is as follows:

$$\psi(t) = \frac{1}{5}e^t + \frac{2}{5}e^{4t} + \frac{2}{5}e^{8t} \quad \text{for } -\infty < t < \infty.$$

Find the probability distribution of X. *Hint:* It is a simple discrete distribution.

12. Suppose that X is a random variable for which the m.g.f. is as follows:

$$\psi(t) = \frac{1}{6}(4 + e^t + e^{-t}) \quad \text{for } -\infty < t < \infty.$$

Find the probability distribution of X.

13. Let X have the Cauchy distribution (see Example 4.1.8). Prove that the m.g.f. $\psi(t)$ is finite only for $t = 0$.

14. Let X have p.d.f.

$$f(x) = \begin{cases} x^{-2} & \text{if } x > 1, \\ 0 & \text{otherwise.} \end{cases}$$

Prove that the m.g.f. $\psi(t)$ is finite for all $t \le 0$ but for no $t > 0$.

15. Prove the following extension of Theorem 4.4.1: If $E(|X|^a) < \infty$ for some positive number a, then $E(|X|^b) < \infty$ for every positive number $b < a$. Give the proof for the case in which X has a discrete distribution.

16. Let X have the binomial distribution with parameters n and p. Let Y have the binomial distribution with parameters n and $1 - p$. Prove that the skewness of Y is the negative of the skewness of X. *Hint:* Let $Z = n - X$ and show that Z has the same distribution as Y.

17. Find the skewness of the distribution in Example 4.4.3.

4.5 The Mean and the Median

Although the mean of a distribution is a measure of central location, the median (see Definition 3.3.3) is also a measure of central location for a distribution. This section presents some comparisons and contrasts between these two location summaries of a distribution.

The Median

It was mentioned in Sec. 4.1 that the mean of a probability distribution on the real line will be at the center of gravity of that distribution. In this sense, the mean of a distribution can be regarded as the *center* of the distribution. There is another point on the line that might also be regarded as the center of the distribution. Suppose that there is a point m_0 that divides the total probability into two equal parts, that is, the probability to the left of m_0 is 1/2, and the probability to the right of m_0 is also 1/2. For a continuous distribution, the median of the distribution introduced in Definition 3.3.3 is such a number. If there is such an m_0, it could legitimately be called a center of the distribution. It should be noted, however, that for some discrete distributions there will not be any point at which the total probability is divided into two parts that are exactly equal. Moreover, for other distributions, which may be either discrete or continuous, there will be more than one such point. Therefore, the formal definition of a median, which will now be given, must be general enough to include these possibilities.

Definition 4.5.1 Median. Let X be a random variable. Every number m with the following property is called a *median* of the distribution of X:

$$\Pr(X \le m) \ge 1/2 \quad \text{and} \quad \Pr(X \ge m) \ge 1/2.$$

Another way to understand this definition is that a median is a point m that satisfies the following two requirements: First, if m is included with the values of X to the left of m, then

$$\Pr(X \le m) \ge \Pr(X > m).$$

Second, if m is included with the values of X to the right of m, then

$$\Pr(X \ge m) \ge \Pr(X < m).$$

If there is a number m such that $\Pr(X < m) = \Pr(X > m)$, that is, if the number m does actually divide the total probability into two equal parts, then m will of course be a median of the distribution of X (see Exercise 16).

Note: Multiple Medians. One can prove that every distribution must have at least one median. Indeed, the 1/2 quantile from Definition 3.3.2 is a median. (See Exercise 1.) For some distributions, every number in some interval is a median. In such

cases, the 1/2 quantile is the minimum of the set of all medians. When a whole interval of numbers are medians of a distribution, some writers refer to the midpoint of the interval as the median.

Example 4.5.1 **The Median of a Discrete Distribution.** Suppose that X has the following discrete distribution:

$$\Pr(X = 1) = 0.1, \qquad \Pr(X = 2) = 0.2,$$
$$\Pr(X = 3) = 0.3, \qquad \Pr(X = 4) = 0.4.$$

The value 3 is a median of this distribution because $\Pr(X \le 3) = 0.6$, which is greater than 1/2, and $\Pr(X \ge 3) = 0.7$, which is also greater than 1/2. Furthermore, 3 is the unique median of this distribution. ◂

Example 4.5.2 **A Discrete Distribution for Which the Median Is Not Unique.** Suppose that X has the following discrete distribution:

$$\Pr(X = 1) = 0.1, \qquad \Pr(X = 2) = 0.4,$$
$$\Pr(X = 3) = 0.3, \qquad \Pr(X = 4) = 0.2.$$

Here, $\Pr(X \le 2) = 1/2$, and $\Pr(X \ge 3) = 1/2$. Therefore, every value of m in the closed interval $2 \le m \le 3$ will be a median of this distribution. The most popular choice of median of this distribution would be the midpoint 2.5. ◂

Example 4.5.3 **The Median of a Continuous Distribution.** Suppose that X has a continuous distribution for which the p.d.f. is as follows:

$$f(x) = \begin{cases} 4x^3 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The unique median of this distribution will be the number m such that

$$\int_0^m 4x^3\, dx = \int_m^1 4x^3\, dx = \frac{1}{2}.$$

This number is $m = 1/2^{1/4}$. ◂

Example 4.5.4 **A Continuous Distribution for Which the Median Is Not Unique.** Suppose that X has a continuous distribution for which the p.d.f. is as follows:

$$f(x) = \begin{cases} 1/2 & \text{for } 0 \le x \le 1, \\ 1 & \text{for } 2.5 \le x \le 3, \\ 0 & \text{otherwise.} \end{cases}$$

Here, for every value of m in the closed interval $1 \le m \le 2.5$, $\Pr(X \le m) = \Pr(X \ge m) = 1/2$. Therefore, every value of m in the interval $1 \le m \le 2.5$ is a median of this distribution. ◂

Comparison of the Mean and the Median

Example 4.5.5 **Last Lottery Number.** In a state lottery game, a three-digit number from 000 to 999 is drawn each day. After several years, all but one of the 1000 possible numbers has been drawn. A lottery official would like to predict how much longer it will be until that missing number is finally drawn. Let X be the number of days ($X = 1$ being tomorrow) until that number appears. It is not difficult to determine the distribution of X, assuming that all 1000 numbers are equally likely to be drawn each day and

that the draws are independent. Let A_x stand for the event that the missing number is drawn on day x for $x = 1, 2, \ldots$. Then $\{X = 1\} = A_1$, and for $x > 1$,

$$\{X = x\} = A_1^c \cap \cdots \cap A_{x-1}^c \cap A_x.$$

Since the A_x events are independent and all have probability 0.001, it is easy to see that the p.f. of X is

$$f(x) = \begin{cases} 0.001(0.999)^{x-1} & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

But, the lottery official wants to give a single-number prediction for when the number will be drawn. What summary of the distribution would be appropriate for this prediction? ◀

The lottery official in Example 4.5.5 wants some sort of "average" or "middle" number to summarize the distribution of the number of days until the last number appears. Presumably she wants a prediction that is neither excessively large nor too small. Either the mean or a median of X can be used as such a summary of the distribution. Some important properties of the mean have already been described in this chapter, and several more properties will be given later in the book. However, for many purposes the median is a more useful measure of the middle of the distribution than is the mean. For example, every distribution has a median, but not every distribution has a mean. As illustrated in Example 4.3.5, the mean of a distribution can be made very large by removing a small but positive amount of probability from any part of the distribution and assigning this amount to a sufficiently large value of x. On the other hand, the median may be unaffected by a similar change in probabilities. If any amount of probability is removed from a value of x larger than the median and assigned to an arbitrarily large value of x, the median of the new distribution will be the same as that of the original distribution. In Example 4.3.5, all numbers in the interval $[0, 1]$ are medians of both random variables X and Y despite the large difference in their means.

Example 4.5.6 Annual Incomes. Suppose that the mean annual income among the families in a certain community is \$30,000. It is possible that only a few families in the community actually have an income as large as \$30,000, but those few families have incomes that are very much larger than \$30,000. As an extreme example, suppose that there are 100 families and 99 of them have income of \$1,000 while the other one has income of \$2,901,000. If, however, the median annual income among the families is \$30,000, then at least one-half of the families must have incomes of \$30,000 or more. ◀

The median has one convenient property that the mean *does not* have.

Theorem 4.5.1 One-to-One Function. Let X be a random variable that takes values in an interval I of real numbers. Let r be a one-to-one function defined on the interval I. If m is a median of X, then $r(m)$ is a median of $r(X)$.

Proof Let $Y = r(X)$. We need to show that $\Pr(Y \geq r(m)) \geq 1/2$ and $\Pr(Y \leq r(m)) \geq 1/2$. Since r is one-to-one on the interval I, it must be either increasing or decreasing over the interval I. If r is increasing, then $Y \geq r(m)$ if and only if $X \geq m$, so $\Pr(Y \geq r(m)) = \Pr(X \geq m) \geq 1/2$. Similarly, $Y \leq r(m)$ if and only if $X \leq m$ and $\Pr(Y \leq r(m)) \geq 1/2$ also. If r is decreasing, then $Y \geq r(m)$ if and only if $X \leq m$. The remainder of the proof is then similar to the preceding. ■

We shall now consider two specific criteria by which the prediction of a random variable X can be judged. By the first criterion, the optimal prediction that can be made is the mean. By the second criterion, the optimal prediction is the median.

Minimizing the Mean Squared Error

Suppose that X is a random variable with mean μ and variance σ^2. Suppose also that the value of X is to be observed in some experiment, but this value must be predicted before the observation can be made. One basis for making the prediction is to select some number d for which the expected value of the square of the error $X - d$ will be a minimum.

Definition 4.5.2 Mean Squared Error/M.S.E.. The number $E[(X - d)^2]$ is called the *mean squared error* (M.S.E.) of the prediction d.

The next result shows that the number d for which the M.S.E. is minimized is $E(X)$.

Theorem 4.5.2 Let X be a random variable with finite variance σ^2, and let $\mu = E(X)$. For every number d,

$$E[(X - \mu)^2] \le E[(X - d)^2]. \tag{4.5.1}$$

Furthermore, there will be equality in the relation (4.5.1) if and only if $d = \mu$.

Proof For every value of d,

$$\begin{aligned} E[(X - d)^2] &= E(X^2 - 2\,dX + d^2) \\ &= E(X^2) - 2\,d\mu + d^2. \end{aligned} \tag{4.5.2}$$

The final expression in Eq. (4.5.2) is simply a quadratic function of d. By elementary differentiation it will be found that the minimum value of this function is attained when $d = \mu$. Hence, in order to minimize the M.S.E., the predicted value of X should be its mean μ. Furthermore, when this prediction is used, the M.S.E. is simply $E[(X - \mu)^2] = \sigma^2$. ∎

Example 4.5.7 Last Lottery Number. In Example 4.5.5, we discussed a state lottery in which one number had never yet been drawn. Let X stand for the number of days until that last number is eventually drawn. The p.f. of X was computed in Example 4.5.5 as

$$f(x) = \begin{cases} 0.001(0.999)^{x-1} & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

We can compute the mean of X as

$$E(X) = \sum_{x=1}^{\infty} x0.001(0.999)^{x-1} = 0.001 \sum_{x=1}^{\infty} x(0.999)^{x-1}. \tag{4.5.3}$$

At first, this sum does not look like one that is easy to compute. However, it is closely related to the general sum

$$g(y) = \sum_{x=0}^{\infty} y^x = \frac{1}{1-y},$$

if $0 < y < 1$. Using properties of power series from calculus, we know that the derivative of $g(y)$ can be found by differentiating the individual terms of the power series. That is,

$$g'(y) = \sum_{x=0}^{\infty} x y^{x-1} = \sum_{x=1}^{\infty} x y^{x-1},$$

for $0 < y < 1$. But we also know that $g'(y) = 1/(1-y)^2$. The last sum in Eq. (4.5.3) is $g'(0.999) = 1/(0.001)^2$. It follows that

$$E(X) = 0.001 \frac{1}{(0.001)^2} = 1000. \quad ◂$$

Minimizing the Mean Absolute Error

Another possible basis for predicting the value of a random variable X is to choose some number d for which $E(|X - d|)$ will be a minimum.

Definition 4.5.3 Mean Absolute Error/M.A.E. The number $E(|X - d|)$ is called the *mean absolute error* (M.A.E.) of the prediction d.

We shall now show that the M.A.E. is minimized when the chosen value of d is a median of the distribution of X.

Theorem 4.5.3 Let X be a random variable with finite mean, and let m be a median of the distribution of X. For every number d,

$$E(|X - m|) \le E(|X - d|). \tag{4.5.4}$$

Furthermore, there will be equality in the relation (4.5.4) if and only if d is also a median of the distribution of X.

Proof For convenience, we shall assume that X has a continuous distribution for which the p.d.f. is f. The proof for any other type of distribution is similar. Suppose first that $d > m$. Then

$$\begin{aligned}
E(|X - d|) - E(|X - m|) &= \int_{-\infty}^{\infty} (|x - d| - |x - m|) f(x)\, dx \\
&= \int_{-\infty}^{m} (d - m) f(x)\, dx + \int_{m}^{d} (d + m - 2x) f(x)\, dx + \int_{d}^{\infty} (m - d) f(x)\, dx \\
&\ge \int_{-\infty}^{m} (d - m) f(x)\, dx + \int_{m}^{d} (m - d) f(x)\, dx + \int_{d}^{\infty} (m - d) f(x)\, dx \\
&= (d - m)[\Pr(X \le m) - \Pr(X > m)]. \qquad (4.5.5)
\end{aligned}$$

Since m is a median of the distribution of X, it follows that

$$\Pr(X \le m) \ge 1/2 \ge \Pr(X > m). \tag{4.5.6}$$

The final difference in the relation (4.5.5) is therefore nonnegative. Hence,

$$E(|X - d|) \ge E(|X - m|). \tag{4.5.7}$$

Furthermore, there can be equality in the relation (4.5.7) only if the inequalities in relations (4.5.5) and (4.5.6) are actually equalities. A careful analysis shows that these inequalities will be equalities only if d is also a median of the distribution of X.

The proof for every value of d such that $d < m$ is similar. ■

Example 4.5.8

Last Lottery Number. In Example 4.5.5, in order to compute the median of X, we must find the smallest number x such that the c.d.f. $F(x) \geq 0.5$. For integer x, we have

$$F(x) = \sum_{n=1}^{x} 0.001(0.999)^{n-1}.$$

We can use the popular formula

$$\sum_{n=0}^{x} y^n = \frac{1 - y^{x+1}}{1 - y}$$

to see that, for integer $x \geq 1$,

$$F(x) = 0.001 \frac{1 - (0.999)^x}{1 - 0.999} = 1 - (0.999)^x.$$

Setting this equal to 0.5 and solving for x gives $x = 692.8$; hence, the median of X is 693. The median is unique because $F(x)$ never takes the exact value 0.5 for any integer x. The median of X is much smaller than the mean of 1000 found in Example 4.5.7. ◀

The reason that the mean is so much larger than the median in Examples 4.5.7 and 4.5.8 is that the distribution has probability at arbitrarily large values but is bounded below. The probability at these large values pulls the mean up because there is no probability at equally small values to balance. The median is not affected by how the upper half of the probability is distributed. The following example involves a symmetric distribution. Here, the mean and median(s) are more similar.

Example 4.5.9

Predicting a Discrete Uniform Random Variable. Suppose that the probability is 1/6 that a random variable X will take each of the following six values: 1, 2, 3, 4, 5, 6. We shall determine the prediction for which the M.S.E. is minimum and the prediction for which the M.A.E. is minimum.

In this example,

$$E(X) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5.$$

Therefore, the M.S.E. will be minimized by the unique value $d = 3.5$.

Also, every number m in the closed interval $3 \leq m \leq 4$ is a median of the given distribution. Therefore, the M.A.E. will be minimized by every value of d such that $3 \leq d \leq 4$ and only by such a value of d. Because the distribution of X is symmetric, the mean of X is also a median of X. ◀

Note: When the M.A.E. and M.S.E. Are Finite. We noted that the median exists for every distribution, but the M.A.E. is finite if and only if the distribution has a finite mean. Similarly, the M.S.E. is finite if and only if the distribution has a finite variance.

Summary

A median of X is any number m such that $\Pr(X \leq m) \geq 1/2$ and $\Pr(X \geq m) \geq 1/2$. To minimize $E(|X - d|)$ by choice of d, one must choose d to be a median of X. To minimize $E[(X - d)^2]$ by choice of d, one must choose $d = E(X)$.

Exercises

1. Prove that the 1/2 quantile as defined in Definition 3.3.2 is a median as defined in Definition 4.5.1.

2. Suppose that a random variable X has a discrete distribution for which the p.f. is as follows:

$$f(x) = \begin{cases} cx & \text{for } x = 1, 2, 3, 4, 5, 6, \\ 0 & \text{otherwise.} \end{cases}$$

Determine all the medians of this distribution.

3. Suppose that a random variable X has a continuous distribution for which the p.d.f. is as follows:

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Determine all the medians of this distribution.

4. In a small community consisting of 153 families, the number of families that have k children ($k = 0, 1, 2, \dots$) is given in the following table:

Number of children	Number of families
0	21
1	40
2	42
3	27
4 or more	23

Determine the mean and the median of the number of children per family. (For the mean, assume that all families with four or more children have only four children. Why doesn't this point matter for the median?)

5. Suppose that an observed value of X is equally likely to come from a continuous distribution for which the p.d.f. is f or from one for which the p.d.f. is g. Suppose that $f(x) > 0$ for $0 < x < 1$ and $f(x) = 0$ otherwise, and suppose also that $g(x) > 0$ for $2 < x < 4$ and $g(x) = 0$ otherwise. Determine: **(a)** the mean and **(b)** the median of the distribution of X.

6. Suppose that a random variable X has a continuous distribution for which the p.d.f. f is as follows:

$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the value of d that minimizes **(a)** $E[(X-d)^2]$ and **(b)** $E(|X-d|)$.

7. Suppose that a person's score X on a certain examination will be a number in the interval $0 \le X \le 1$ and that X has a continuous distribution for which the p.d.f. is as follows:

$$f(x) = \begin{cases} x + \frac{1}{2} & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the prediction of X that minimizes **(a)** the M.S.E. and **(b)** the M.A.E.

8. Suppose that the distribution of a random variable X is symmetric with respect to the point $x = 0$ and that $E(X^4) < \infty$. Show that $E[(X-d)^4]$ is minimized by the value $d = 0$.

9. Suppose that a fire can occur at any one of five points along a road. These points are located at -3, -1, 0, 1, and 2 in Fig. 4.9. Suppose also that the probability that each of these points will be the location of the next fire that occurs along the road is as specified in Fig. 4.9.

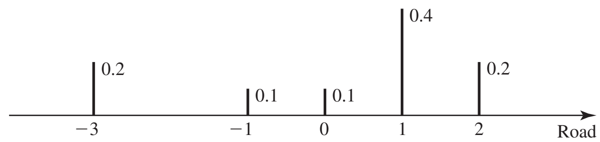


Figure 4.9 Probabilities for Exercise 9.

- **a.** At what point along the road should a fire engine wait in order to minimize the expected value of the square of the distance that it must travel to the next fire?
- **b.** Where should the fire engine wait to minimize the expected value of the distance that it must travel to the next fire?

10. If n houses are located at various points along a straight road, at what point along the road should a store be located in order to minimize the sum of the distances from the n houses to the store?

11. Let X be a random variable having the binomial distribution with parameters $n = 7$ and $p = 1/4$, and let Y be a random variable having the binomial distribution with parameters $n = 5$ and $p = 1/2$. Which of these two random variables can be predicted with the smaller M.S.E.?

12. Consider a coin for which the probability of obtaining a head on each given toss is 0.3. Suppose that the coin is to be tossed 15 times, and let X denote the number of heads that will be obtained.

- **a.** What prediction of X has the smallest M.S.E.?
- **b.** What prediction of X has the smallest M.A.E.?

13. Suppose that the distribution of X is symmetric around a point m. Prove that m is a median of X.

14. Find the median of the Cauchy distribution defined in Example 4.1.8.

15. Let X be a random variable with c.d.f. F. Suppose that $a < b$ are numbers such that both a and b are medians of X.

a. Prove that $F(a) = 1/2$.

b. Prove that there exist a smallest $c \le a$ and a largest $d \ge b$ such that every number in the closed interval $[c, d]$ is a median of X.

c. If X has a discrete distribution, prove that $F(d) > 1/2$.

16. Let X be a random variable. Suppose that there exists a number m such that $\Pr(X < m) = \Pr(X > m)$. Prove that m is a median of the distribution of X.

17. Let X be a random variable. Suppose that there exists a number m such that $\Pr(X < m) < 1/2$ and $\Pr(X > m) < 1/2$. Prove that m is the unique median of the distribution of X.

18. Prove the following extension of Theorem 4.5.1. Let m be the p quantile of the random variable X. (See Definition 3.3.2.) If r is a strictly increasing function, then $r(m)$ is the p quantile of $r(X)$.

4.6 Covariance and Correlation

When we are interested in the joint distribution of two random variables, it is useful to have a summary of how much the two random variables depend on each other. The covariance and correlation are attempts to measure that dependence, but they only capture a particular type of dependence, namely linear dependence.

Covariance

Example 4.6.1 Test Scores. When applying for college, high school students often take a number of standardized tests. Consider a particular student who will take both a verbal and a quantitative test. Let X be this student's score on the verbal test, and let Y be the same student's score on the quantitative test. Although there are students who do much better on one test than the other, it might still be reasonable to expect that a student who does very well on one test to do at least a little better than average on the other. We would like to find a numerical summary of the joint distribution of X and Y that reflects the degree to which we believe a high or low score on one test will be accompanied by a high or low score on the other test. ◀

When we consider the joint distribution of two random variables, the means, the medians, and the variances of the variables provide useful information about their marginal distributions. However, these values do not provide any information about the relationship between the two variables or about their tendency to vary together rather than independently. In this section and the next one, we shall introduce summaries of a joint distribution that enable us to measure the association between two random variables, determine the variance of the sum of an arbitrary number of dependent random variables, and predict the value of one random variable by using the observed value of some other related variable.

Definition 4.6.1 Covariance. Let X and Y be random variables having finite means. Let $E(X) = \mu_X$ and $E(Y) = \mu_Y$ The *covariance of X and Y*, which is denoted by $\text{Cov}(X, Y)$, is defined as

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)], \tag{4.6.1}$$

if the expectation in Eq. (4.6.1) exists.

It can be shown (see Exercise 2 at the end of this section) that if both X and Y have finite variance, then the expectation in Eq. (4.6.1) will exist and $\text{Cov}(X, Y)$ will be finite. However, the value of $\text{Cov}(X, Y)$ can be positive, negative, or zero.

Example 4.6.2 **Test Scores.** Let X and Y be the test scores in Example 4.6.1, and suppose that they have the joint p.d.f.

$$f(x, y) = \begin{cases} 2xy + 0.5 & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We shall compute the covariance $\text{Cov}(X, Y)$. First, we shall compute the means μ_X and μ_Y of X and Y, respectively. The symmetry in the joint p.d.f. means that X and Y have the same marginal distribution; hence, $\mu_X = \mu_Y$. We see that

$$\begin{aligned} \mu_X &= \int_0^1 \int_0^1 [2x^2 y + 0.5x] dy dx \\ &= \int_0^1 [x^2 + 0.5x] dx = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}, \end{aligned}$$

so that $\mu_Y = 7/12$ as well. The covariance can be computed using Theorem 4.1.2. Specifically, we must evaluate the integral

$$\int_0^1 \int_0^1 \left(x - \frac{7}{12}\right)\left(y - \frac{7}{12}\right)(2xy + 0.5)\, dy\, dx.$$

This integral is straightforward, albeit tedious, to compute, and the result is $\text{Cov}(X, Y) = 1/144$. ◄

The following result often simplifies the calculation of a covariance.

Theorem 4.6.1 For all random variables X and Y such that $\sigma_X^2 < \infty$ and $\sigma_Y^2 < \infty$,

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y). \tag{4.6.2}$$

Proof It follows from Eq. (4.6.1) that

$$\begin{aligned} \text{Cov}(X, Y) &= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y. \end{aligned}$$

Since $E(X) = \mu_X$ and $E(Y) = \mu_Y$, Eq. (4.6.2) is obtained. ■

The covariance between X and Y is intended to measure the degree to which X and Y tend to be large at the same time or the degree to which one tends to be large while the other is small. Some intution about this interpretation can be gathered from a careful look at Eq. (4.6.1). For example, suppose that $\text{Cov}(X, Y)$ is positive. Then $X > \mu_X$ and $Y > \mu_Y$ must occur together and/or $X < \mu_X$ and $Y < \mu_Y$ must occur together to a larger extent than $X < \mu_X$ occurs with $Y > \mu_Y$ and $X > \mu_X$ occurs with $Y < \mu_Y$. Otherwise, the mean would be negative. Similarly, if $\text{Cov}(X, Y)$ is negative, then $X > \mu_X$ and $Y < \mu_Y$ must occur together and/or $X < \mu_X$ and $Y > \mu_Y$ must occur together to larger extent than the other two inequalities. If $\text{Cov}(X, Y) = 0$, then the extent to which X and Y are on the same sides of their respective means exactly balances the extent to which they are on opposite sides of their means.

Correlation

Although Cov(X, Y) gives a numerical measure of the degree to which X and Y vary together, the magnitude of Cov(X, Y) is also influenced by the overall magnitudes of X and Y. For example, in Exercise 5 in this section, you can prove that $\text{Cov}(2X, Y) = 2\,\text{Cov}(X, Y)$. In order to obtain a measure of association between X and Y that is *not driven by arbitrary changes in the scales* of one or the other random variable, we define a slightly different quantity next.

Definition 4.6.2 Correlation. Let X and Y be random variables with finite variances σ_X^2 and σ_Y^2, respectively. Then the *correlation of X and Y*, which is denoted by $\rho(X, Y)$, is defined as follows:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{4.6.3}$$

In order to determine the range of possible values of the correlation $\rho(X, Y)$, we shall need the following result.

Theorem 4.6.2 Schwarz Inequality. For all random variables U and V such that $E(UV)$ exists,

$$[E(UV)]^2 \le E(U^2)E(V^2). \tag{4.6.4}$$

If, in addition, the right-hand side of Eq. (4.6.4) is finite, then the two sides of Eq. (4.6.4) equal the same value if and only if there are nonzero constants a and b such that $aU + bV = 0$ with probability 1.

Proof If $E(U^2) = 0$, then $\Pr(U = 0) = 1$. Therefore, it must also be true that $\Pr(UV = 0) = 1$. Hence, $E(UV) = 0$, and the relation (4.6.4) is satisfied. Similarly, if $E(V^2) = 0$, then the relation (4.6.4) will be satisfied. Moreover, if either $E(U^2)$ or $E(V^2)$ is infinite, then the right side of the relation (4.6.4) will be infinite. In this case, the relation (4.6.4) will surely be satisfied.

For the rest of the proof, assume that $0 < E(U^2) < \infty$ and $0 < E(V^2) < \infty$. For all numbers a and b,

$$0 \le E[(aU + bV)^2] = a^2E(U^2) + b^2E(V^2) + 2abE(UV) \tag{4.6.5}$$

and

$$0 \le E[(aU - bV)^2] = a^2E(U^2) + b^2E(V^2) - 2abE(UV). \tag{4.6.6}$$

If we let $a = [E(V^2)]^{1/2}$ and $b = [E(U^2)]^{1/2}$, then it follows from the relation (4.6.5) that

$$E(UV) \ge -[E(U^2)E(V^2)]^{1/2}. \tag{4.6.7}$$

It also follows from the relation (4.6.6) that

$$E(UV) \le [E(U^2)E(V^2)]^{1/2}. \tag{4.6.8}$$

These two relations together imply that the relation (4.6.4) is satisfied.

Finally, suppose that the right-hand side of Eq. (4.6.4) is finite. Both sides of (4.6.4) equal the same value if and only if the same is true for either (4.6.7) or (4.6.8). Both sides of (4.6.7) equal the same value if and only if the rightmost expression in (4.6.5) is 0. This, in turn, is true if and only if $E[(aU + bV)^2] = 0$, which occurs if and only if $aU + bV = 0$ with probability 1. The reader can easily check that both sides of (4.6.8) equal the same value if and only if $aU - bV = 0$ with probability 1. ∎

A slight variant on Theorem 4.6.2 is the result we want.

Theorem 4.6.3 Cauchy-Schwarz Inequality. Let X and Y be random variables with finite variance. Then

$$[\mathrm{Cov}(X, Y)]^2 \le \sigma_X^2 \sigma_Y^2, \tag{4.6.9}$$

and

$$-1 \le \rho(X, Y) \le 1. \tag{4.6.10}$$

Furthermore, the inequality in Eq. (4.6.9) is an equality if and only if there are nonzero constants a and b and a constant c such that $aX + bY = c$ with probability 1.

Proof Let $U = X - \mu_X$ and $V = Y - \mu_Y$. Eq. (4.6.9) now follows directly from Theorem 4.6.2. In turn, it follows from Eq. (4.6.3) that $[\rho(X, Y)]^2 \le 1$ or, equivalently, that Eq. (4.6.10) holds. The final claim follows easily from the similar claim at the end of Theorem 4.6.2. ■

Definition 4.6.3 Positively/Negatively Correlated/Uncorrelated. It is said that X and Y are *positively correlated* if $\rho(X, Y) > 0$, that X and Y are *negatively correlated* if $\rho(X, Y) < 0$, and that X and Y are *uncorrelated* if $\rho(X, Y) = 0$.

It can be seen from Eq. (4.6.3) that $\mathrm{Cov}(X, Y)$ and $\rho(X, Y)$ must have the same sign; that is, both are positive, or both are negative, or both are zero.

Example 4.6.3 Test Scores. For the two test scores in Example 4.6.2, we can compute the correlation $\rho(X, Y)$. The variances of X and Y are both equal to 11/144, so the correlation is $\rho(X, Y) = 1/11$. ◀

Properties of Covariance and Correlation

We shall now present four theorems pertaining to the basic properties of covariance and correlation.

The first theorem shows that independent random variables must be uncorrelated.

Theorem 4.6.4 If X and Y are independent random variables with $0 < \sigma_X^2 < \infty$ and $0 < \sigma_Y^2 < \infty$, then

$$\mathrm{Cov}(X, Y) = \rho(X, Y) = 0.$$

Proof If X and Y are independent, then $E(XY) = E(X)E(Y)$. Therefore, by Eq. (4.6.2), $\mathrm{Cov}(X, Y) = 0$. Also, it follows that $\rho(X, Y) = 0$. ■

The converse of Theorem 4.6.4 is not true as a general rule. Two dependent random variables can be uncorrelated. Indeed, even though Y is an explicit function of X, it is possible that $\rho(X, Y) = 0$, as in the following examples.

Example 4.6.4 Dependent but Uncorrelated Random Variables. Suppose that the random variable X can take only the three values -1, 0, and 1, and that each of these three values has the same probability. Also, let the random variable Y be defined by the relation $Y = X^2$. We shall show that X and Y are dependent but uncorrelated.

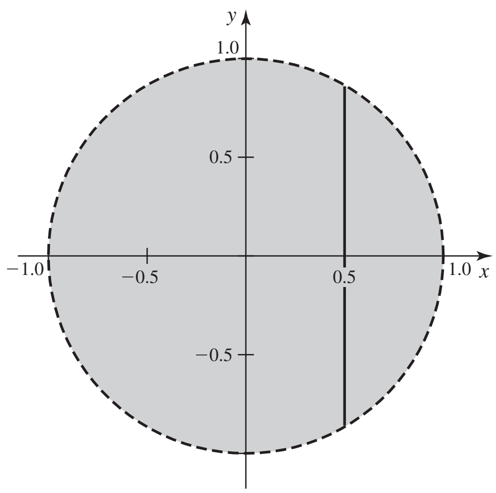


Figure 4.10 The shaded region is where the joint p.d.f. of (X, Y) is constant and nonzero in Example 4.6.5. The vertical line indicates the values of Y that are possible when $X = 0.5$.

In this example, X and Y are clearly dependent, since Y is not constant and the value of Y is completely determined by the value of X. However,

$$E(XY) = E(X^3) = E(X) = 0,$$

because X^3 is the same random variable as X. Since $E(XY) = 0$ and $E(X)E(Y) = 0$, it follows from Theorem 4.6.1 that $\text{Cov}(X, Y) = 0$ and that X and Y are uncorrelated. ◀

Example 4.6.5 Uniform Distribution Inside a Circle. Let (X, Y) have joint p.d.f. that is constant on the interior of the unit circle, the shaded region in Fig. 4.10. The constant value of the p.d.f. is one over the area of the circle, that is, $1/(2\pi)$. It is clear that X and Y are dependent since the region where the joint p.d.f. is nonzero is not a rectangle. In particular, notice that the set of possible values for Y is the interval (−1, 1), but when $X = 0.5$, the set of possible values for Y is the smaller interval (−0.866, 0.866). The symmetry of the circle makes it clear that both X and Y have mean 0. Also, it is not difficult to see that $E(XY) = \int\int xyf(x, y)dxdy = 0$. To see this, notice that the integral of xy over the top half of the circle is exactly the negative of the integral of xy over the bottom half. Hence, $\text{Cov}(X, Y) = 0$, but the random variables are dependent. ◀

The next result shows that if Y is a *linear* function of X, then X and Y must be correlated and, in fact, $|\rho(X, Y)| = 1$.

Theorem 4.6.5 Suppose that X is a random variable such that $0 < \sigma_X^2 < \infty$, and $Y = aX + b$ for some constants a and b, where $a \neq 0$. If $a > 0$, then $\rho(X, Y) = 1$. If $a < 0$, then $\rho(X, Y) = -1$.

Proof If $Y = aX + b$, then $\mu_Y = a\mu_X + b$ and $Y - \mu_Y = a(X - \mu_X)$. Therefore, by Eq. (4.6.1),

$$\text{Cov}(X, Y) = aE[(X - \mu_X)^2] = a\sigma_X^2.$$

Since $\sigma_Y = |a|\sigma_X$, the theorem follows from Eq. (4.6.3). ■

There is a converse to Theorem 4.6.5. That is, $|\rho(X, Y)| = 1$ implies that X and Y are linearly related. (See Exercise 17.) In general, the value of $\rho(X, Y)$ provides a measure of the extent to which two random variables X and Y are linearly related. If

the joint distribution of X and Y is relatively concentrated around a straight line in the xy-plane that has a positive slope, then $\rho(X, Y)$ will typically be close to 1. If the joint distribution is relatively concentrated around a straight line that has a negative slope, then $\rho(X, Y)$ will typically be close to -1. We shall not discuss these concepts further here, but we shall consider them again when the bivariate normal distribution is introduced and studied in Sec. 5.10.

Note: Correlation Measures Only Linear Relationship. A large value of $|\rho(X, Y)|$ means that X and Y are close to being linearly related and hence are closely related. But a small value of $|\rho(X, Y)|$ does not mean that X and Y are not close to being related. Indeed, Example 4.6.4 illustrates random variables that are functionally related but have 0 correlation.

We shall now determine the variance of the sum of random variables that are not necessarily independent.

Theorem 4.6.6 If X and Y are random variables such that $\text{Var}(X) < \infty$ and $\text{Var}(Y) < \infty$, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y). \tag{4.6.11}$$

Proof Since $E(X + Y) = \mu_X + \mu_Y$, then

$$\begin{aligned}\text{Var}(X + Y) &= E[(X + Y - \mu_X - \mu_Y)^2]\\ &= E[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)]\\ &= \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y).\end{aligned}$$

■

For all constants a and b, it can be shown that $\text{Cov}(aX, bY) = ab\,\text{Cov}(X, Y)$ (see Exercise 5 at the end of this section). The following then follows easily from Theorem 4.6.6.

Corollary 4.6.1 Let a, b, and c be constants. Under the conditions of Theorem 4.6.6,

$$\text{Var}(aX + bY + c) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\,\text{Cov}(X, Y). \tag{4.6.12}$$

■

A particularly useful special case of Corollary 4.6.1 is

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\,\text{Cov}(X, Y). \tag{4.6.13}$$

Example 4.6.6 Investment Portfolio. Consider, once again, the investor in Example 4.3.7 on page 230 trying to choose a portfolio with \$100,000 to invest. We shall make the same assumptions about the returns on the two stocks, except that now we will suppose that the correlation between the two returns R_1 and R_2 is -0.3, reflecting a belief that the two stocks tend to react in opposite ways to common market forces. The variance of a portfolio of s_1 shares of the first stock, s_2 shares of the second stock, and s_3 dollars invested at 3.6% is now

$$\text{Var}(s_1R_1 + s_2R_2 + 0.036s_3) = 55s_1^2 + 28s_2^2 - 0.3\sqrt{55 \times 28}s_1s_2.$$

We continue to assume that (4.3.2) holds. Figure 4.11 shows the relationship between the mean and variance of the efficient portfolios in this example and Example 4.3.7. Notice how the variances are smaller in this example than in Example 4.3.7. This is due to the fact that the negative correlation lowers the variance of a linear combination with positive coefficients. ◄

Theorem 4.6.6 can also be extended easily to the variance of the sum of n random variables, as follows.

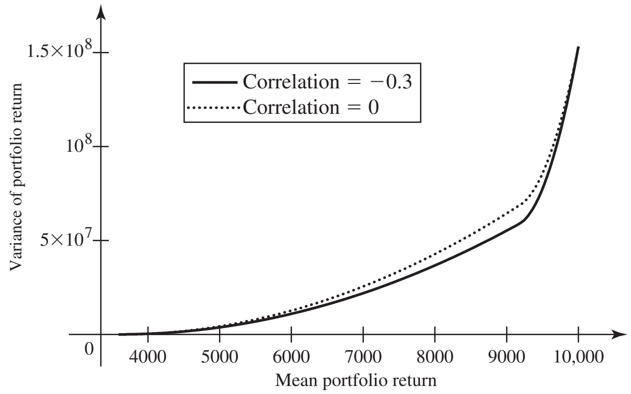


Figure 4.11 Mean and variance of efficient investment portfolios.

Theorem 4.6.7 If $X_1, \dots, X_n$ are random variables such that $\text{Var}(X_i) < \infty$ for $i = 1, \dots, n$, then

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum\sum_{i<j} \text{Cov}(X_i, X_j). \tag{4.6.14}$$

Proof For every random variable Y, $\text{Cov}(Y, Y) = \text{Var}(Y)$. Therefore, by using the result in Exercise 8 at the end of this section, we can obtain the following relation:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j).$$

We shall separate the final sum in this relation into two sums: (i) the sum of those terms for which $i = j$ and (ii) the sum of those terms for which $i \neq j$. Then, if we use the fact that $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$, we obtain the relation

$$\begin{aligned}\text{Var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \text{Var}(X_i) + \sum\sum_{i \neq j} \text{Cov}(X_i, X_j)\\ &= \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum\sum_{i<j} \text{Cov}(X_i, X_j).\end{aligned}$$

■

The following is a simple corrolary to Theorem 4.6.7.

Corollary 4.6.2 If $X_1, \dots, X_n$ are uncorrelated random variables (that is, if X_i and X_j are uncorrelated whenever $i \neq j$), then

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i). \tag{4.6.15}$$

■

Corollary 4.6.2 extends Theorem 4.3.5 on page 230, which states that (4.6.15) holds if $X_1, \dots, X_n$ are independent random variables.

Note: In General, Variances Add Only for Uncorrelated Random Variables. The variance of a sum of random variables should be calculated using Theorem 4.6.7 in general. Corollary 4.6.2 applies only for uncorrelated random variables.

Summary

The covariance of X and Y is $\text{Cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$. The correlation is $\rho(X, Y) = \text{Cov}(X, Y)/[\text{Var}(X)\,\text{Var}(Y)]^{1/2}$, and it measures the extent to which X and Y are linearly related. Indeed, X and Y are precisely linearly related if and only if $|\rho(X, Y)| = 1$. The variance of a sum of random variables can be expressed as the sum of the variances plus two times the sum of the covariances. The variance of a linear function is $\text{Var}(aX + bY + c) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\,\text{Cov}(X, Y)$.

Exercises

1. Suppose that the pair (X, Y) is uniformly distributed on the interior of a circle of radius 1. Compute $\rho(X, Y)$.

2. Prove that if $\text{Var}(X) < \infty$ and $\text{Var}(Y) < \infty$, then $\text{Cov}(X, Y)$ is finite. *Hint:* By considering the relation $[(X - \mu_X) \pm (Y - \mu_Y)]^2 \ge 0$, show that

$$|(X - \mu_X)(Y - \mu_Y)| \le \frac{1}{2}[(X - \mu_X)^2 + (Y - \mu_Y)^2].$$

3. Suppose that X has the uniform distribution on the interval $[-2, 2]$ and $Y = X^6$. Show that X and Y are uncorrelated.

4. Suppose that the distribution of a random variable X is symmetric with respect to the point $x = 0$, $0 < E(X^4) < \infty$, and $Y = X^2$. Show that X and Y are uncorrelated.

5. For all random variables X and Y and all constants a, b, c, and d, show that

$$\text{Cov}(aX + b, cY + d) = ac\,\text{Cov}(X, Y).$$

6. Let X and Y be random variables such that $0 < \sigma_X^2 < \infty$ and $0 < \sigma_Y^2 < \infty$. Suppose that $U = aX + b$ and $V = cY + d$, where $a \neq 0$ and $c \neq 0$. Show that $\rho(U, V) = \rho(X, Y)$ if $ac > 0$, and $\rho(U, V) = -\rho(X, Y)$ if $ac < 0$.

7. Let X, Y, and Z be three random variables such that $\text{Cov}(X, Z)$ and $\text{Cov}(Y, Z)$ exist, and let a, b, and c be arbitrary given constants. Show that

$$\text{Cov}(aX + bY + c, Z) = a\,\text{Cov}(X, Z) + b\,\text{Cov}(Y, Z).$$

8. Suppose that $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ are random variables such that $\text{Cov}(X_i, Y_j)$ exists for $i = 1, \dots, m$ and $j = 1, \dots, n$, and suppose that $a_1, \dots, a_m$ and $b_1, \dots, b_n$ are constants. Show that

$$\text{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j\,\text{Cov}(X_i, Y_j).$$

9. Suppose that X and Y are two random variables, which may be dependent, and $\text{Var}(X) = \text{Var}(Y)$. Assuming that $0 < \text{Var}(X + Y) < \infty$ and $0 < \text{Var}(X - Y) < \infty$, show that the random variables $X + Y$ and $X - Y$ are uncorrelated.

10. Suppose that X and Y are negatively correlated. Is $\text{Var}(X + Y)$ larger or smaller than $\text{Var}(X - Y)$?

11. Show that two random variables X and Y cannot possibly have the following properties: $E(X) = 3$, $E(Y) = 2$, $E(X^2) = 10$, $E(Y^2) = 29$, and $E(XY) = 0$.

12. Suppose that X and Y have a continuous joint distribution for which the joint p.d.f. is as follows:

$$f(x, y) = \begin{cases} \frac{1}{3}(x + y) & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the value of $\text{Var}(2X - 3Y + 8)$.

13. Suppose that X and Y are random variables such that $\text{Var}(X) = 9$, $\text{Var}(Y) = 4$, and $\rho(X, Y) = -1/6$. Determine **(a)** $\text{Var}(X + Y)$ and **(b)** $\text{Var}(X - 3Y + 4)$.

14. Suppose that X, Y, and Z are three random variables such that $\text{Var}(X) = 1$, $\text{Var}(Y) = 4$, $\text{Var}(Z) = 8$, $\text{Cov}(X, Y) = 1$, $\text{Cov}(X, Z) = -1$, and $\text{Cov}(Y, Z) = 2$. Determine **(a)** $\text{Var}(X + Y + Z)$ and **(b)** $\text{Var}(3X - Y - 2Z + 1)$.

15. Suppose that $X_1, \dots, X_n$ are random variables such that the variance of each variable is 1 and the correlation between each pair of different variables is 1/4. Determine $\text{Var}(X_1 + \cdots + X_n)$.

16. Consider the investor in Example 4.2.3 on page 220. Suppose that the returns R_1 and R_2 on the two stocks have correlation -1. A portfolio will consist of s_1 shares of the first stock and s_2 shares of the second stock where $s_1, s_2 \ge 0$. Find a portfolio such that the total cost of the portfolio is \$6000 and the variance of the return is 0. Why is this situation unrealistic?

17. Let X and Y be random variables with finite variance. Prove that $|\rho(X, Y)| = 1$ implies that there exist constants a, b, and c such that $aX + bY = c$ with probability 1. *Hint:* Use Theorem 4.6.2 with $U = X - \mu_X$ and $V = Y - \mu_Y$.

18. Let X and Y have a continuous distribution with joint p.d.f.

$$f(x, y) = \begin{cases} x + y & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Compute the covariance $\text{Cov}(X, Y)$.

4.7 Conditional Expectation

Since expectations (including variances and covariances) are properties of distributions, there will exist conditional versions of all such distributional summaries as well as conditional versions of all theorems that we have proven or will later prove about expectations. In particular, suppose that we wish to predict one random variable Y using a function $d(X)$ of another random variable X so as to minimize $E([Y - d(X)]^2)$. Then $d(X)$ should be the conditional mean of Y given X. There is also a very useful theorem that is an extension to expectations of the law of total probability.

Definition and Basic Properties

Example 4.7.1 Household Survey. A collection of households were surveyed, and each household reported the number of members and the number of automobiles owned. The reported numbers are in Table 4.1.

Suppose that we were to sample a household at random from those households in the survey and learn the number of members. What would then be the expected number of automobiles that they own? ◀

The question at the end of Example 4.7.1 is closely related to the conditional distribution of one random variable given the other, as defined in Sec. 3.6.

Definition 4.7.1 Conditional Expectation/Mean. Let X and Y be random variables such that the mean of Y exists and is finite. The *conditional expectation (or conditional mean) of Y given $X = x$* is denoted by $E(Y|x)$ and is defined to be the expectation of the conditional distribution of Y given $X = x$.

For example, if Y has a continuous conditional distribution given $X = x$ with conditional p.d.f. $g_2(y|x)$, then

$$E(Y|x) = \int_{-\infty}^{\infty} y g_2(y|x)\, dy. \tag{4.7.1}$$

Similarly, if Y has a discrete conditional distribution given $X = x$ with conditional p.f. $g_2(y|x)$, then

$$E(Y|x) = \sum_{\text{All } y} y g_2(y|x). \tag{4.7.2}$$

Table 4.1 Reported numbers of household members and automobiles in Example 4.7.1

Number of automobiles	Number of members: 1	2	3	4	5	6	7	8
0	10	7	3	2	2	1	0	0
1	12	21	25	30	25	15	5	1
2	1	5	10	15	20	11	5	3
3	0	2	3	5	5	3	2	1

The value of $E(Y|x)$ will not be uniquely defined for those values of x such that the marginal p.f. or p.d.f. of X satisfies $f_1(x) = 0$. However, since these values of x form a set of points whose probability is 0, the definition of $E(Y|x)$ at such a point is irrelevant. (See Exercise 11 in Sec. 3.6.) It is also possible that there will be some values of x such that the mean of the conditional distribution of Y given $X = x$ is undefined for those x values. When the mean of Y exists and is finite, the set of x values for which the conditional mean is undefined has probability 0.

The expressions in Eqs. (4.7.1) and (4.7.2) are functions of x. These functions of x can be computed before X is observed, and this idea leads to the following useful concept.

Definition 4.7.2 **Conditional Means as Random Variables.** Let $h(x)$ stand for the function of x that is denoted $E(Y|x)$ in either (4.7.1) or (4.7.2). Define the symbol $E(Y|X)$ to mean $h(X)$ and call it the *conditional mean of Y given X*.

In other words, $E(Y|X)$ is a random variable (a function of X) whose value when $X = x$ is $E(Y|x)$. Obviously, we could define $E(X|Y)$ and $E(X|y)$ analogously.

Example 4.7.2 **Household Survey.** Consider the household survey in Example 4.7.1. Let X be the number of members in a randomly selected household from the survey, and let Y be the number of cars owned by that household. The 250 surveyed households are all equally likely to be selected, so $\Pr(X = x, Y = y)$ is the number of households with x members and y cars, divided by 250. Those probabilities are reported in Table 4.2. Suppose that the sampled household has $X = 4$ members. The conditional p.f. of Y given $X = 4$ is $g_2(y|4) = f(4, y)/f_1(4)$, which is the $x = 4$ column of Table 4.2 divided by $f_1(4) = 0.208$, namely,

$$g_2(0|4) = 0.0385, \quad g_2(1|4) = 0.5769, \quad g_2(2|4) = 0.2885, \quad g_2(3|4) = 0.0962.$$

The conditional mean of Y given $X = 4$ is then

$$E(Y|4) = 0 \times 0.0385 + 1 \times 0.5769 + 2 \times 0.2885 + 3 \times 0.0962 = 1.442.$$

Similarly, we can compute $E(Y|x)$ for all eight values of x. They are

x	1	2	3	4	5	6	7	8
$E(Y\|x)$	0.609	1.057	1.317	1.442	1.538	1.533	1.75	2

Table 4.2 Joint p.f. $f(x, y)$ of X and Y in Example 4.7.2 together with marginal p.f.'s $f_1(x)$ and $f_2(y)$

	x								
y	1	2	3	4	5	6	7	8	$f_2(y)$
0	0.040	0.028	0.012	0.008	0.008	0.004	0	0	0.100
1	0.048	0.084	0.100	0.120	0.100	0.060	0.020	0.004	0.536
2	0.004	0.020	0.040	0.060	0.080	0.044	0.020	0.012	0.280
3	0	0.008	0.012	0.020	0.020	0.012	0.008	0.004	0.084
$f_1(x)$	0.092	0.140	0.164	0.208	0.208	0.120	0.048	0.020	

The random variable that takes the value 0.609 when the sampled household has one member, takes the value 1.057 when the sampled household has two members, and so on, is the random variable $E(Y|X)$. ◀

Example 4.7.3 A Clinical Trial. Consider a clinical trial in which a number of patients will be treated and each patient will have one of two possible outcomes: success or failure. Let P be the proportion of successes in a very large collection of patients, and let $X_i = 1$ if the ith patient is a success and $X_i = 0$ if not. Assume that the random variables $X_1, X_2, \ldots$ are conditionally independent given $P = p$ with $\Pr(X_i = 1|P = p) = p$. Let $X = X_1 + \cdots + X_n$, which is the number of patients out of the first n who are successes. We now compute the conditional mean of X given P. The patients are independent and identically distributed conditional on $P = p$. Hence, the conditional distribution of X given $P = p$ is the binomial distribution with parameters n and p. As we saw in Sec. 4.2, the mean of this binomial distribution is np, so $E(X|p) = np$ and $E(X|P) = nP$. Later, we will show how to compute the conditional mean of P given X. This can be used to predict P after observing X. ◀

Note: The Conditional Mean of Y Given X Is a Random Variable. Because $E(Y|X)$ is a function of the random variable X, it is itself a random variable with its own probability distribution, which can be derived from the distribution of X. On the other hand, $h(x) = E(Y|x)$ is a function of x that can be manipulated like any other function. The connection between the two is that when one substitutes the random variable X for x in $h(x)$, the result is $h(X) = E(Y|X)$.

We shall now show that the mean of the random variable $E(Y|X)$ must be $E(Y)$. A similar calculation shows that the mean of $E(X|Y)$ must be $E(X)$.

Theorem 4.7.1 Law of Total Probability for Expectations. Let X and Y be random variables such that Y has finite mean. Then

$$E[E(Y|X)] = E(Y). \tag{4.7.3}$$

Proof We shall assume, for convenience, that X and Y have a continuous joint distribution. Then

$$\begin{aligned} E[E(Y|X)] &= \int_{-\infty}^{\infty} E(Y|x) f_1(x)\, dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y g_2(y|x) f_1(x)\, dy\, dx. \end{aligned}$$

Since $g_2(y|x) = f(x, y)/f_1(x)$, it follows that

$$E[E(Y|X)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y)\, dy\, dx = E(Y).$$

The proof for a discrete distribution or a more general type of distribution is similar. ■

Example 4.7.4 Household Survey. At the end of Example 4.7.2, we described the random variable $E(Y|X)$. Its distribution can be constructed from that description. It has a discrete distribution that takes the eight values of $E(Y|x)$ listed near the end of that example with corresponding probabilities $f_1(x)$ for $x = 1, \ldots, 8$. To be specific, let $Z = E(Y|X)$, then $\Pr[Z = E(Y|x)] = f_1(x)$ for $x = 1, \ldots, 8$. The specific values are

z	0.609	1.057	1.317	1.442	1.538	1.533	1.75	2
$\Pr(Z = z)$	0.092	0.140	0.164	0.208	0.208	0.120	0.048	0.020

We can compute $E(Z) = 0.609 \times 0.092 + \cdots + 2 \times 0.020 = 1.348$. The reader can verify that $E(Y) = 1.348$ by using the values of $f_2(y)$ in Table 4.2. ◀

Example 4.7.5 A Clinical Trial. In Example 4.7.3, we let X be the number of patients out of the first n who are successes. The conditional mean of X given $P = p$ was computed as $E(X|p) = np$, where P is the proportion of successes in a large population of patients. If the distribution of P is uniform on the interval [0, 1], then the marginal expected value of X is $E[E(X|P)] = E(nP) = n/2$. We will see how to calculate $E(P|X)$ in Example 4.7.8. ◀

Example 4.7.6 Choosing Points from Uniform Distributions. Suppose that a point X is chosen in accordance with the uniform distribution on the interval [0, 1]. Also, suppose that after the value $X = x$ has been observed $(0 < x < 1)$, a point Y is chosen in accordance with a uniform distribution on the interval $[x, 1]$. We shall determine the value of $E(Y)$.

For each given value of x $(0 < x < 1)$, $E(Y|x)$ will be equal to the midpoint $(1/2)(x + 1)$ of the interval $[x, 1]$. Therefore, $E(Y|X) = (1/2)(X + 1)$ and

$$E(Y) = E[E(Y|X)] = \frac{1}{2}[E(X) + 1] = \frac{1}{2}\left(\frac{1}{2} + 1\right) = \frac{3}{4}.$$ ◀

When manipulating the conditional distribution given $X = x$, it is safe to act as if X is the constant x. This fact, which can simplify the calculation of certain conditional means, is now stated without proof.

Theorem 4.7.2 Let X and Y be random variables, and let $Z = r(X, Y)$ for some function r. The conditional distribution of Z given $X = x$ is the same as the conditional distribution of $r(x, Y)$ given $X = x$. ■

One consequence of Theorem 4.7.2 when X and Y have a continuous joint distribution is that

$$E(Z|x) = E(r(x, Y)|x) = \int_{-\infty}^{\infty} r(x, y) g_2(y|x)\, dy.$$

Theorem 4.7.1 also implies that for two arbitrary random variables X and Y,

$$E\{E[r(X, Y)|X]\} = E[r(X, Y)], \tag{4.7.4}$$

by letting $Z = r(X, Y)$ and noting that $E\{E(Z|X)\} = E(Z)$.

We can define, in a similar manner, the conditional expectation of $r(X, Y)$ given Y and the conditional expectation of a function $r(X_1, \ldots, X_n)$ of several random variables given one or more of the variables $X_1, \ldots, X_n$.

Example 4.7.7 Linear Conditional Expectation. Suppose that $E(Y|X) = aX + b$ for some constants a and b. We shall determine the value of $E(XY)$ in terms of $E(X)$ and $E(X^2)$.

By Eq. (4.7.4), $E(XY) = E[E(XY|X)]$. Furthermore, since X is considered to be given and fixed in the conditional expectation,

$$E(XY|X) = XE(Y|X) = X(aX + b) = aX^2 + bX.$$

Therefore,

$$E(XY) = E(aX^2 + bX) = aE(X^2) + bE(X). \qquad \blacktriangleleft$$

The mean is not the only feature of a conditional distribution that is important enough to get its own name.

Definition 4.7.3 Conditional Variance. For every given value x, let $\text{Var}(Y|x)$ denote the variance of the conditional distribution of Y given that $X = x$. That is,

$$\text{Var}(Y|x) = E\{[Y - E(Y|x)]^2|x\}. \tag{4.7.5}$$

We call $\text{Var}(Y|x)$ the *conditional variance of* Y *given* $X = x$.

The expression in Eq. (4.7.5) is once again a function $v(x)$. We shall define $\text{Var}(Y|X)$ to be $v(X)$ and call it the *conditional variance of* Y *given* X.

Note: Other Conditional Quantities. In much the same way as in Definitions 4.7.1 and 4.7.3, we could define any conditional summary of a distribution that we wish. For example, conditional quantiles of Y given $X = x$ are the quantiles of the conditional distribution of Y given $X = x$. The conditional m.g.f. of Y given $X = x$ is the m.g.f. of the conditional distribution of Y given $X = x$, etc.

Prediction

At the end of Example 4.7.3, we considered the problem of predicting the proportion P of successes in a large population of patients given the observed number X of succeses in a sample of size n. In general, consider two arbitrary random variables X and Y that have a specified joint distribution and suppose that after the value of X has been observed, the value of Y must be predicted. In other words, the predicted value of Y can depend on the value of X. We shall assume that this predicted value $d(X)$ must be chosen so as to minimize the mean squared error $E\{[Y - d(X)]^2\}$.

Theorem 4.7.3 The prediction $d(X)$ that minimizes $E\{[Y - d(X)]^2\}$ is $d(X) = E(Y|X)$.

Proof We shall prove the theorem in the case in which X has a continuous distribution, but the proof in the discrete case is virtually identical. Let $d(X) = E(Y|X)$, and let $d^*(X)$ be an arbitrary predictor. We need only prove that $E\{[Y - d(X)]^2\} \le E\{[Y - d^*(X)]^2\}$. It follows from Eq. (4.7.4) that

$$E\{[Y - d(X)]^2\} = E(E\{[Y - d(X)]^2|X\}). \tag{4.7.6}$$

A similar equation holds for d^*. Let $Z = [Y - d(X)]^2$, and let $h(x) = E(Z|x)$. Similarly, let $Z^* = [Y - d^*(X)]^2$ and $h^*(x) = E(Z^*|x)$. The right-hand side of (4.7.6) is $\int h(x) f_1(x)\, dx$, and the corresponding expression using d^* is $\int h^*(x) f_1(x)\, dx$. So, the proof will be complete if we can prove that

$$\int h(x) f_1(x)\, dx \le \int h^*(x) f_1(x)\, dx. \tag{4.7.7}$$

Clearly, Eq. (4.7.7) holds if we can show that $h(x) \le h^*(x)$ for all x. That is, the proof is complete if we can show that $E\{[Y - d(X)]^2|x\} \le E\{[Y - d^*(X)]^2|x\}$. When we condition on $X = x$, we are allowed to treat X as if it were the constant x, so we need to show that $E\{[Y - d(x)]^2|x\} \le E\{[Y - d^*(x)]^2|x\}$. These last expressions are nothing more than the M.S.E.'s for two different predictions $d(x)$ and $d^*(x)$ of Y calculated

using the conditional distribution of Y given $X = x$. As discussed in Sec. 4.5, the M.S.E. of such a prediction is smallest if the prediction is the mean of the distribution of Y. In this case, that mean is the mean of the conditional distribution of Y given $X = x$. Since $d(x)$ is the mean of the conditional distribution of Y given $X = x$, it must have smaller M.S.E. than every other prediction $d^*(x)$. Hence, $h(x) \le h^*(x)$ for all x. ■

If the value $X = x$ is observed and the value $E(Y|x)$ is predicted for Y, then the M.S.E. of this predicted value will be $\text{Var}(Y|x)$, from Definition 4.7.3. It follows from Eq. (4.7.6) that if the prediction is to be made by using the function $d(X) = E(Y|X)$, then the overall M.S.E., averaged over all the possible values of X, will be $E[\text{Var}(Y|X)]$.

If the value of Y must be predicted without any information about the value of X, then, as shown in Sec. 4.5, the best prediction is the mean $E(Y)$ and the M.S.E. is $\text{Var}(Y)$. However, if X can be observed before the prediction is made, the best prediction is $d(X) = E(Y|X)$ and the M.S.E. is $E[\text{Var}(Y|X)]$. Thus, the reduction in the M.S.E. that can be achieved by using the observation X is

$$\text{Var}(Y) - E[\text{Var}(Y|X)]. \tag{4.7.8}$$

This reduction provides a measure of the usefulness of X in predicting Y. It is shown in Exercise 11 at the end of this section that this reduction can also be expressed as $\text{Var}[E(Y|X)]$.

It is important to distinguish carefully between the overall M.S.E., which is $E[\text{Var}(Y|X)]$, and the M.S.E. of the particular prediction to be made when $X = x$, which is $\text{Var}(Y|x)$. *Before* the value of X has been observed, the appropriate value for the M.S.E. of the complete process of observing X and then predicting Y is $E[\text{Var}(Y|X)]$. *After* a particular value x of X has been observed and the prediction $E(Y|x)$ has been made, the appropriate measure of the M.S.E. of this prediction is $\text{Var}(Y|x)$. A useful relationship between these values is given in the following result, whose proof is left to Exercise 11.

Theorem 4.7.4 Law of Total Probability for Variances. If X and Y are arbitrary random variables for which the necessary expectations and variances exist, then $\text{Var}(Y) = E[\text{Var}(Y|X)] + \text{Var}[E(Y|X)]$. ■

Example 4.7.8 A Clinical Trial. In Example 4.7.3, let X be the number of patients out of the first 40 in a clinical trial who have success as their outcome. Let P be the probability that an individual patient is a success. Suppose that P has the uniform distribution on the interval [0, 1] before the trial begins, and suppose that the outcomes of the patients are conditionally independent given $P = p$. As we saw in Example 4.7.3, X has the binomial distribution with parameters 40 and p given $P = p$. If we needed to minimize M.S.E. in predicting P before observing X, we would use the mean of P, namely, 1/2. The M.S.E. would be $\text{Var}(P) = 1/12$. However, we shall soon observe the value of X and then predict P. To do this, we shall need the conditional distribution of P given $X = x$. Bayes' theorem for random variables (3.6.13) tells us that the conditional p.d.f. of P given $X = x$ is

$$g_2(p|x) = \frac{g_1(x|p) f_2(p)}{f_1(x)}, \tag{4.7.9}$$

where $g_1(x|p)$ is the conditional p.f. of X given $P = p$, namely, the binomial p.f. $g_1(x|p) = \binom{40}{x} p^x (1-p)^{40-x}$ for $x = 0, \ldots, 40$, $f_2(p) = 1$ for $0 < p < 1$ is the marginal p.d.f. of P, and $f_1(x)$ is the marginal p.f. of X obtained from the law of total probability

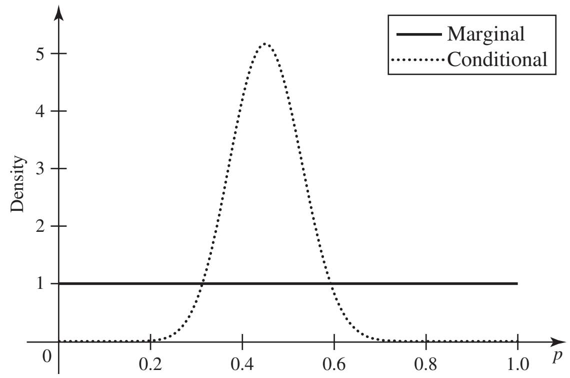


Figure 4.12 The conditional p.d.f. of P given $X = 18$ in Example 4.7.8. The marginal p.d.f. of P (prior to observing X) is also shown.

for random variables (3.6.12):

$$f_1(x) = \int_0^1 \binom{40}{x} p^x (1-p)^{40-x}\, dp. \tag{4.7.10}$$

This last integral looks difficult to compute. However, there is a simple formula for integrals of this form, namely,

$$\int_0^1 p^k (1-p)^\ell \, dp = \frac{k!\ell!}{(k+\ell+1)!}. \tag{4.7.11}$$

A proof of Eq. (4.7.11) is given in Sec. 5.8. Substituting (4.7.11) into (4.7.10) yields

$$f_1(x) = \frac{40!}{x!(40-x)!}\frac{x!(40-x)!}{41!} = \frac{1}{41},$$

for $x = 0, \dots, 40$. Substituting this into Eq. (4.7.9) yields

$$g_2(p|x) = \frac{41!}{x!(40-x)!} p^x (1-p)^{40-x}, \quad \text{for } 0 < p < 1.$$

For example, with $x = 18$, the observed number of successes in Table 2.1, a graph of $g_2(p|18)$ is shown in Fig. 4.12.

If we want to minimize the M.S.E. when predicting P, we should use $E(P|x)$, the conditional mean. We can compute $E(P|x)$ using the conditional p.d.f. and Eq. (4.7.11):

$$\begin{aligned} E(P|x) &= \int_0^1 p \frac{41!}{x!(40-x)!} p^x (1-p)^{40-x}\, dp \\ &= \frac{41!}{x!(40-x)!}\frac{(x+1)!(40-x)!}{42!} = \frac{x+1}{42}. \end{aligned} \tag{4.7.12}$$

So, after $X = x$ is observed, we will predict P to be $(x+1)/42$, which is very close to the proportion of the first 40 patients who are successes. The M.S.E. after observing $X = x$ is the conditional variance $\text{Var}(P|x)$. We can compute this using (4.7.12) and

$$\begin{aligned} E(P^2|x) &= \int_0^1 p^2 \frac{41!}{x!(40-x)!} p^x (1-p)^{40-x}\, dp \\ &= \frac{41!}{x!(40-x)!}\frac{(x+2)!(40-x)!}{43!} = \frac{(x+1)(x+2)}{42 \times 43}. \end{aligned}$$

Using the fact that $\text{Var}(P|x) = E(P^2|x) - [E(P|x)]^2$, we see that

$$\text{Var}(P|x) = \frac{(x+1)(41-x)}{42^2 \times 43}.$$

The overall M.S.E. of predicting P from X is the mean of the conditional M.S.E.

$$\begin{aligned}
E[\text{Var}(P|X)] &= E\left(\frac{(X+1)(41-X)}{42^2 \times 43}\right) \\
&= \frac{1}{75{,}852} E(-X^2 + 40X + 41) \\
&= \frac{1}{75{,}852}\left(-\frac{1}{41}\sum_{x=0}^{40} x^2 + \frac{40}{41}\sum_{x=0}^{40} x + 41\right) \\
&= \frac{1}{75{,}852}\left(-\frac{1}{41}\frac{40 \times 41 \times 81}{6} + \frac{40}{41}\frac{40 \times 41}{2} + 41\right) \\
&= \frac{301}{75{,}852} = 0.003968.
\end{aligned}$$

In this calculation, we used two popular formulas,

$$\sum_{k=0}^{n} k = \frac{n(n+1)}{2}, \tag{4.7.13}$$

$$\sum_{k=0}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}. \tag{4.7.14}$$

The overall M.S.E. is quite a bit smaller than the value $1/12 = 0.08333$, which we would have obtained before observing X. As an illustration, Fig. 4.12 shows how much more spread out the marginal distribution of P is compared to the conditional distribution of P after observing $X = 18$. ◀

It should be emphasized that for the conditions of Example 4.7.8, 0.003968 is the appropriate value of the overall M.S.E. when it is known that the value of X will be available for predicting P but before the explicit value of X has been determined. After the value of $X = x$ has been determined, the appropriate value of the M.S.E. is $\text{Var}(P|x) = \frac{(x+1)(41-x)}{75{,}852}$. Notice that the largest possible value of $\text{Var}(P|x)$ is 0.005814 when $x = 20$ and is still much less than 1/12.

A result similar to Theorem 4.7.3 holds if we are trying to minimize the M.A.E. (mean absolute error) of our prediction rather than the M.S.E. In Exercise 16, you can prove that the predictor that minimizes M.A.E. is $d(X)$ equal to the median of the conditional distribution of Y given X.

Summary

The conditional mean $E(Y|x)$ of Y given $X = x$ is the mean of the conditional distribution of Y given $X = x$. This conditional distribution was defined in Chapter 3. Likewise, the conditional variance $\text{Var}(Y|x)$ of Y given $X = x$ is the variance of the conditional distribution. The law of total probability for expectations says that $E[E(Y|X)] = E(Y)$. If we will observe X and then need to predict Y, the predictor that leads to the smallest M.S.E. is the conditional mean $E(Y|X)$.

Exercises

1. Consider again the situation described in Example 4.7.8. Compute the M.S.E. when using $E(P|x)$ to predict P after observing $X = 18$. How much smaller is this than the marginal M.S.E. 1/12?

2. Suppose that 20 percent of the students who took a certain test were from school A and that the arithmetic average of their scores on the test was 80. Suppose also that 30 percent of the students were from school B and that the arithmetic average of their scores was 76. Suppose, finally, that the other 50 percent of the students were from school C and that the arithmetic average of their scores was 84. If a student is selected at random from the entire group that took the test, what is the expected value of her score?

3. Suppose that $0 < \text{Var}(X) < \infty$ and $0 < \text{Var}(Y) < \infty$. Show that if $E(X|Y)$ is constant for all values of Y, then X and Y are uncorrelated.

4. Suppose that the distribution of X is symmetric with respect to the point $x = 0$, that all moments of X exist, and that $E(Y|X) = aX + b$, where a and b are given constants. Show that X^{2m} and Y are uncorrelated for $m = 1, 2, \ldots$.

5. Suppose that a point X_1 is chosen from the uniform distribution on the interval [0, 1], and that after the value $X_1 = x_1$ is observed, a point X_2 is chosen from a uniform distribution on the interval $[x_1, 1]$. Suppose further that additional variables $X_3, X_4, \ldots$ are generated in the same way. In general, for $j = 1, 2, \ldots$, after the value $X_j = x_j$ has been observed, X_{j+1} is chosen from a uniform distribution on the interval $[x_j, 1]$. Find the value of $E(X_n)$.

6. Suppose that the joint distribution of X and Y is the uniform distribution on the circle $x^2 + y^2 < 1$. Find $E(X|Y)$.

7. Suppose that X and Y have a continuous joint distribution for which the joint p.d.f. is as follows:

$$f(x, y) = \begin{cases} x + y & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find $E(Y|X)$ and $\text{Var}(Y|X)$.

8. Consider again the conditions of Exercise 7. **(a)** If it is observed that $X = 1/2$, what predicted value of Y will have the smallest M.S.E.? **(b)** What will be the value of this M.S.E.?

9. Consider again the conditions of Exercise 7. If the value of Y is to be predicted from the value of X, what will be the minimum value of the overall M.S.E.?

10. Suppose that, for the conditions in Exercises 7 and 9, a person either can pay a cost c for the opportunity of observing the value of X before predicting the value of Y or can simply predict the value of Y without first observing the value of X. If the person considers her total loss to be the cost c plus the M.S.E. of her predicted value, what is the maximum value of c that she should be willing to pay?

11. Prove Theorem 4.7.4.

12. Suppose that X and Y are random variables such that $E(Y|X) = aX + b$. Assuming that $\text{Cov}(X, Y)$ exists and that $0 < \text{Var}(X) < \infty$, determine expressions for a and b in terms of $E(X)$, $E(Y)$, $\text{Var}(X)$, and $\text{Cov}(X, Y)$.

13. Suppose that a person's score X on a mathematics aptitude test is a number in the interval (0, 1) and that his score Y on a music aptitude test is also a number in the interval (0, 1). Suppose also that in the population of all college students in the United States, the scores X and Y are distributed in accordance with the following joint p.d.f.:

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y) & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

- **a.** If a college student is selected at random, what predicted value of his score on the music test has the smallest M.S.E.?
- **b.** What predicted value of his score on the mathematics test has the smallest M.A.E.?

14. Consider again the conditions of Exercise 13. Are the scores of college students on the mathematics test and the music test positively correlated, negatively correlated, or uncorrelated?

15. Consider again the conditions of Exercise 13. **(a)** If a student's score on the mathematics test is 0.8, what predicted value of his score on the music test has the smallest M.S.E.? **(b)** If a student's score on the music test is 1/3, what predicted value of his score on the mathematics test has the smallest M.A.E.?

16. Define a conditional median of Y given $X = x$ to be any median of the conditional distribution of Y given $X = x$. Suppose that we will get to observe X and then we will need to predict Y. Suppose that we wish to choose our prediction $d(X)$ so as to minimize mean absolute error, $E(|Y - d(X)|)$. Prove that $d(x)$ should be chosen to be a conditional median of Y given $X = x$. *Hint:* You can modify the proof of Theorem 4.7.3 to handle this case.

17. Prove Theorem 4.7.2 for the case in which X and Y have a discrete joint distribution. The key to the proof is to write all of the necessary conditional p.f.'s in terms of the joint p.f. of X and Y and the marginal p.f. of X. To facilitate this, for each x and z, give a name to the set of y values such that $r(x, y) = z$.

★ 4.8 Utility

Much of statistical inference consists of choosing between several available actions. Generally, we do not know for certain which choice will be best, because some important random variable has not yet been observed. For some values of that random variable one choice is best, and for other values some other choice is best. We can try to weigh the costs and benefits of the various choices against the probabilities that the various choices turn out to be best. Utility is one tool for assigning values to the costs and benefits of our choices. The expected value of the utility then balances the costs and benefits according to how likely the uncertain possibilities are.

Utility Functions

Example 4.8.1 Choice of Gambles. Consider two gambles between which a gambler must choose. Each gamble will be expressed as a random variable for which positive values mean a gain to the gambler and negative values mean a loss to the gambler. The numerical values of each random variable tell the number of dollars that the gambler gains or loses. Let X have the p.f.

$$f(x) = \begin{cases} 0.5 & \text{if } x = 500 \text{ or } x = -350, \\ 0 & \text{otherwise,} \end{cases}$$

and let Y have the p.f.

$$g(y) = \begin{cases} 1/3 & \text{if } y = 40,\ y = 50, \text{ or } y = 60, \\ 0 & \text{otherwise,} \end{cases}$$

It is simple to compute that $E(X) = 75$ and $E(Y) = 50$. How might a gambler choose between these two gambles? Is X better than Y simply because it has higher expected value? ◀

In Example 4.8.1, a gambler who does not desire to risk losing 350 dollars for the chance of winning 500 dollars might prefer Y, which yields a certain gain of at least 40 dollars.

The *theory of utility* was developed during the 1930s and 1940s to describe a person's preference among gambles like those in Example 4.8.1. According to that theory, a person will prefer a gamble X for which the expectation of a certain function $U(X)$ is a maximum, rather than a gamble for which simply the expected gain $E(X)$ is a maximum.

Definition 4.8.1 Utility Function. A person's *utility function* U is a function that assigns to each possible amount x $(-\infty < x < \infty)$ a number $U(x)$ representing the actual worth to the person of gaining the amount x.

Example 4.8.2 Choice of Gambles. Suppose that a person's utility function is U and that she must choose between the gambles X and Y in Example 4.8.1. Then

$$E[U(X)] = \frac{1}{2}U(500) + \frac{1}{2}U(-350) \tag{4.8.1}$$

and

$$E[U(Y)] = \frac{1}{3}U(60) + \frac{1}{3}U(50) + \frac{1}{3}U(40). \tag{4.8.2}$$

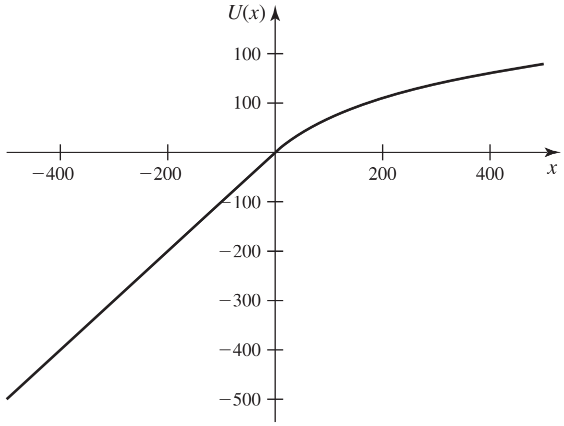


Figure 4.13 The utility function for Example 4.8.2.

The person would prefer the gamble for which the expected utility of the gain, as specified by Eq. (4.8.1) or Eq. (4.8.2), is larger.

As a specific example, consider the following utility function that penalizes losses to a much greater extent than it rewards gains:

$$U(x) = \begin{cases} 100\log(x + 100) - 461 & \text{if } x \ge 0, \\ x & \text{if } x < 0. \end{cases} \tag{4.8.3}$$

This function was chosen to be differentiable at $x = 0$, continuous everywhere, increasing, concave for $x > 0$, and linear for $x < 0$. A graph of $U(x)$ is given in Fig. 4.13. Using this specific U, we compute

$$E[U(X)] = \frac{1}{2}[100\log(600) - 461] + \frac{1}{2}(-350) = -85.4,$$

$$\begin{aligned} E[U(Y)] &= \frac{1}{3}[100\log(160) - 461] + \frac{1}{3}[100\log(150) - 461] + \frac{1}{3}[100\log(140) - 461] \\ &= 40.4. \end{aligned}$$

We see that a person with the utility function in Eq. (4.8.3) would prefer Y to X. ◀

Here, we formalize the principle that underlies the choice between gambles illustrated in Example 4.8.1.

Definition 4.8.2 Maximizing Expected Utility. We say that a person chooses between gambles by *maximizing expected utility* if the following conditions hold. There is a utility function U, and when the person must choose between any two gambles X and Y, he will prefer X to Y if $E[U(X)] > E[U(Y)]$ and will be indifferent between X and Y if $E[U(X)] = E[U(Y)]$.

In words, Definition 4.8.2 says that a person chooses between gambles by maximizing expected utility if he will choose a gamble X for which $E[U(X)]$ is a maximum.

If one adopts a utility function, then one can (at least in principle) make choices between gambles by maximizing expected utility. The computational algorithms necessary to perform the maximization often provide a practical challenge. Conversely, if one makes choices between gambles in such a way that certain reasonable criteria apply, then one can prove that there exists a utility function such that the choices

correspond to maximizing expected utility. We shall not consider this latter problem in detail here; however, it is discussed by DeGroot (1970) and Schervish (1995, chapter 3) along with other aspects of the theory of utility.

Examples of Utility Functions

Since it is reasonable to assume that every person prefers a larger gain to a smaller gain, we shall assume that every utility function $U(x)$ is an increasing function of the gain x. However, the shape of the function $U(x)$ will vary from person to person and will depend on each person's willingness to risk losses of various amounts in attempting to increase his gains.

For example, consider two gambles X and Y for which the gains have the following probability distributions:

$$\Pr(X = -3) = 0.5, \quad \Pr(X = 2.5) = 0.4, \quad \Pr(X = 6) = 0.1 \tag{4.8.4}$$

and

$$\Pr(Y = -2) = 0.3, \quad \Pr(Y = 1) = 0.4, \quad \Pr(Y = 3) = 0.3. \tag{4.8.5}$$

We shall assume that a person must choose one of the following three decisions: (i) accept gamble X, (ii) accept gamble Y, or (iii) do not accept either gamble. We shall now determine the decision that a person would choose for three different utility functions.

Example 4.8.3 Linear Utility Function. Suppose that $U(x) = ax + b$ for some constants a and b, where $a > 0$. In this case, for every gamble X, $E[U(X)] = aE(X) + b$. Hence, for every two gambles X and Y, $E[U(X)] > E[U(Y)]$ if and only if $E(X) > E(Y)$. In other words, a person who has a linear utility function will always choose a gamble for which the expected gain is a maximum.

When the gambles X and Y are defined by Eqs. (4.8.4) and (4.8.5),

$$E(X) = (0.5)(-3) + (0.4)(2.5) + (0.1)(6) = 0.1$$

and

$$E(Y) = (0.3)(-2) + (0.4)(1) + (0.3)(3) = 0.7.$$

Furthermore, since the gain from not accepting either of these gambles is 0, the expected gain from choosing not to accept either gamble is clearly 0. Since $E(Y) > E(X) > 0$, it follows that a person who has a linear utility function would choose to accept gamble Y. If gamble Y were not available, then the person would prefer to accept gamble X rather than not to gamble at all. ◀

Example 4.8.4 Cubic Utility Function. Suppose that a person's utility function is $U(x) = x^3$ for $-\infty < x < \infty$. Then for the gambles defined by Eqs. (4.8.4) and (4.8.5),

$$E[U(X)] = (0.5)(-3)^3 + (0.4)(2.5)^3 + (0.1)(6)^3 = 14.35$$

and

$$E[U(Y)] = (0.3)(-2)^3 + (0.4)(1)^3 + (0.3)(3)^3 = 6.1.$$

Furthermore, the utility of not accepting either gamble is $U(0) = 0^3 = 0$. Since $E[U(X)] > E[U(Y)] > 0$, it follows that the person would choose to accept gamble X. If gamble X were not available, the person would prefer to accept gamble Y rather than not to gamble at all. ◀

Example 4.8.5 **Logarithmic Utility Function.** Suppose that a person's utility function is $U(x) = \log(x + 4)$ for $x > -4$. Since $\lim_{x\to -4} \log(x + 4) = -\infty$, a person who has this utility function cannot choose a gamble in which there is any possibility of her gain being -4 or less. For the gambles X and Y defined by Eqs. (4.8.4) and (4.8.5),

$$E[U(X)] = (0.5)(\log 1) + (0.4)(\log 6.5) + (0.1)(\log 10) = 0.9790$$

and

$$E[U(Y)] = (0.3)(\log 2) + (0.4)(\log 5) + (0.3)(\log 7) = 1.4355.$$

Furthermore, the utility of not accepting either gamble is $U(0) = \log 4 = 1.3863$. Since $E[U(Y)] > U(0) > E[U(X)]$, it follows that the person would choose to accept gamble Y. If gamble Y were not available, the person would prefer not to gamble at all rather than to accept gamble X. ◂

Selling a Lottery Ticket

Suppose that a person has a lottery ticket from which she will receive a random gain of X dollars, where X has a specified probability distribution. We shall determine the number of dollars for which the person would be willing to sell this lottery ticket.

Let U denote the person's utility function. Then the expected utility of her gain from the lottery ticket is $E[U(X)]$. If she sells the lottery ticket for x_0 dollars, then her gain is x_0 dollars, and the utility of this gain is $U(x_0)$. The person would prefer to accept x_0 dollars as a certain gain rather than accept the random gain X from the lottery ticket if and only if $U(x_0) > E[U(X)]$. Hence, the person would be willing to sell the lottery ticket for any amount x_0 such that $U(x_0) > E[U(X)]$. If $U(x_0) = E[U(X)]$, she would be equally willing to either sell the lottery ticket or accept the random gain X.

Example 4.8.6 **Quadratic Utility Function.** Suppose that $U(x) = x^2$ for $x \ge 0$, and suppose that the person has a lottery ticket from which she will win either 36 dollars with probability 1/4 or 0 dollars with probability 3/4. For how many dollars x_0 would she be willing to sell this lottery ticket?

The expected utility of the gain from the lottery ticket is

$$E[U(X)] = \frac{1}{4}U(36) + \frac{3}{4}U(0) = \frac{1}{4}(36^2) + \frac{3}{4}(0) = 324.$$

Therefore, the person would be willing to sell the lottery ticket for any amount x_0 such that $U(x_0) = x_0^2 > 324$. Hence, $x_0 > 18$. In other words, although the expected gain from the lottery ticket in this example is only 9 dollars, the person would not sell the ticket for less than 18 dollars. ◂

Example 4.8.7 **Square Root Utility Function.** Suppose now that $U(x) = x^{1/2}$ for $x \ge 0$, and consider again the lottery ticket described in Example 4.8.6. The expected utility of the gain from the lottery ticket in this case is

$$E[U(X)] = \frac{1}{4}U(36) + \frac{3}{4}U(0) = \frac{1}{4}(6) + \frac{3}{4}(0) = 1.5.$$

Therefore, the person would be willing to sell the lottery ticket for any amount x_0 such that $U(x_0) = x_0^{1/2} > 1.5$. Hence, $x_0 > 2.25$. In other words, although the expected gain from the lottery ticket in this example is 9 dollars, the person would be willing to sell the ticket for as little as 2.25 dollars. ◂

Some Statistical Decision Problems

Much of the theory of statistical inference (the subject of Chapters 7–11 of this text) deals with problems in which one has to make one of several available choices. Generally, which choice is best depends on some random variable that has not yet been observed. One example was already discussed in Sec. 4.5, where we introduced the mean squared error (M.S.E.) and mean absolute error (M.A.E.) criteria for predicting a random variable. In these cases, we have to choose a number d for our prediction of a random variable Y. Which prediction will be best depends on the value of Y that we do not yet know. Random variables like $-|Y-d|$ and $-(Y-d)^2$ are gambles, and the choice of gamble that minimizes M.A.E. or M.S.E. is the choice that maximizes an expected utility.

Example 4.8.8 Predicting a Random Variable. Suppose that Y is a random variable that we need to predict. For each possible prediction d, there is a gamble $X_d = -|Y-d|$ that specifies our gain when we are being judged by absolute error. Alternatively, if we are being judged by squared error, the appropriate gamble to consider would be $Z_d = -(Y-d)^2$. Notice that these gambles are always negative, meaning that our gain is negative because we lose according to how far Y is from the prediction d. If our utility U is linear, then maximizing $E[U(X_d)]$ by choice of d is the same as minimizing M.A.E. Also, maximizing $E[U(Z_d)]$ by choice of d is the same as minimizing M.S.E. The equivalence between maximizing expected utility and minimizing the mean error would continue to hold if the prediction were allowed to depend on another random variable W that we could observe before predicting. That is, our prediction would be a function $d(W)$, and $X_d = -|Y-d(W)|$ or $Z_d = -[Y-d(W)]^2$ would be the gamble whose expected utility we would want to compute. ◄

Example 4.8.9 Bounding a Random Variable. Suppose that Y is a random variable and that we are interested in whether or not $Y \le c$ for some constant c. For example, Y could be the random variable P in our clinical trial Example 4.7.3. We might be interested in whether or not $P \le p_0$, where p_0 is the probability that a patient will be a success without any help from the treatment being studied. Suppose that we have to make one of two available decisions:

(t) continue to promote the treatment, or
(a) abandon the treatment.

If we choose t, suppose that we stand to gain

$$X_t = \begin{cases} 10^6 & \text{if } P > p_0, \\ -10^6 & \text{if } P \le p_0. \end{cases}$$

If we choose a, our gain will be $X_a = 0$. If our utility function is U, then the expected utility for choosing t is $E[U(X_t)]$, and t would be the better choice if this value is greater than $U(0)$. For example, suppose that our utility is

$$U(x) = \begin{cases} x^{0.8} & \text{if } x \ge 0, \\ x & \text{if } x < 0. \end{cases} \tag{4.8.6}$$

Then $U(0) = 0$ and

$$\begin{aligned} E[U(X_t)] &= -10^6 \Pr(P \le p_0) + [10^6]^{0.8} \Pr(P > p_0) \\ &= 10^{4.8} - (10^6 + 10^{4.8}) \Pr(P \le p_0). \end{aligned}$$

So, $E[U(X_t)] > 0$ if $\Pr(P \le p_0) < 10^{4.8}/(10^6 + 10^{4.8}) = 0.0594$. It makes sense that t is better than a if $\Pr(P \le p_0)$ is small. The reason is that the utility of choosing t over a is only positive when $P > p_0$. This example is in the spirit of hypothesis testing, which will be the subject of Chapter 9. ◄

Example 4.8.10 **Investment.** In Example 4.2.2, we compared two possible stock purchases based on their expected returns and value at risk, VaR. Suppose that the investor has a nonlinear utility function for dollars. To be specific, suppose that the utility of a return of x would equal $U(x)$ given in Eq. (4.8.6). We can calculate the expected utility of the return from each of the two possible stock purchases in Example 4.2.2 to decide which is more favorable. If R is the return per share and we buy s shares, then the return is $X = sR$, and the expected utility of the return is

$$E[U(sR)] = \int_{-\infty}^{0} srf(r)\,dr + \int_{0}^{\infty} (sr)^{0.8} f(r)\,dr, \tag{4.8.7}$$

where f is the p.d.f. of R. For the first stock, the return per share is R_1 distributed uniformly on the interval $[-10, 20]$, and the number of shares would be $s_1 = 120$. This makes (4.8.7) equal to

$$E[U(120R_1)] = \int_{-10}^{0} \frac{120r}{30}\,dr + \int_{0}^{20} \frac{(120r)^{0.8}}{30}\,dr = -12.6.$$

For the second stock, the return per share is R_2 distributed uniformly on the interval $[-4.5, 10]$, and the number of shares would be $s_2 = 200$. This makes (4.8.7) equal to

$$E[U(200R_2)] = \int_{-4.5}^{0} \frac{200r}{14.5}\,dr + \int_{0}^{10} \frac{(200r)^{0.8}}{14.5}\,dr = 27.9.$$

With this utility function, the expected utility of the first stock purchase is actually negative because the big gains (up to $120 \times 20 = 2400$) add less to the utility ($2400^{0.8} = 506$) than the big losses (up to $120 \times -10 = -1200$) take away from the utility. The second stock purchase has positive expected utility, so it would be the preferred choice in this example. ◄

Summary

When we have to make choices in the face of uncertainty, we need to assess what our gains and losses will be under each of the uncertain possibilities. Utility is the value to us of those gains and losses. For example, if X represents the random gain from a possible choice, then $U(X)$ is the value to us of the random gain we would receive if we were to make that choice. We should make the choice such that $E[U(X)]$ is as large as possible.

Exercises

1. Let $\alpha > 0$. A decision maker has a utility function for money of the form

$$U(x) = \begin{cases} x^{\alpha} & \text{if } x > 0, \\ x & \text{if } x \le 0. \end{cases}$$

Suppose that this decision maker is trying to decide whether or not to buy a lottery ticket for \$1. The lottery ticket pays \$500 with probability 0.001, and it pays \$0 with probability 0.999. What would the values of α have to be in order for this decision maker to prefer buying the ticket to not buying it?

2. Consider three gambles X, Y, and Z for which the probability distributions of the gains are as follows:

$$\Pr(X=5)=\Pr(X=25)=1/2,$$
$$\Pr(Y=10)=\Pr(Y=20)=1/2,$$
$$\Pr(Z=15)=1.$$

Suppose that a person's utility function has the form $U(x)=x^2$ for $x>0$. Which of the three gambles would she prefer?

3. Determine which of the three gambles in Exercise 2 would be preferred by a person whose utility function is $U(x)=x^{1/2}$ for $x>0$.

4. Determine which of the three gambles in Exercise 2 would be preferred by a person whose utility function has the form $U(x)=ax+b$, where a and b are constants $(a>0)$.

5. Consider a utility function U for which $U(0)=0$ and $U(100)=1$. Suppose that a person who has this utility function is indifferent to either accepting a gamble from which his gain will be 0 dollars with probability 1/3 or 100 dollars with probability 2/3 or accepting 50 dollars as a sure thing. What is the value of $U(50)$?

6. Consider a utility function U for which $U(0)=5$, $U(1)=8$, and $U(2)=10$. Suppose that a person who has this utility function is indifferent to either of two gambles X and Y, for which the probability distributions of the gains are as follows:

$$\Pr(X=-1)=0.6,\ \Pr(X=0)=0.2,\ \Pr(X=2)=0.2;$$
$$\Pr(Y=0)=0.9,\quad \Pr(Y=1)=0.1.$$

What is the value of $U(-1)$?

7. Suppose that a person must accept a gamble X of the following form:

$$\Pr(X=a)=p \quad\text{and}\quad \Pr(X=1-a)=1-p,$$

where p is a given number such that $0<p<1$. Suppose also that the person can choose and fix the value of a $(0\le a\le 1)$ to be used in this gamble. Determine the value of a that the person would choose if his utility function was $U(x)=\log x$ for $x>0$.

8. Determine the value of a that a person would choose in Exercise 7 if his utility function was $U(x)=x^{1/2}$ for $x\ge 0$.

9. Determine the value of a that a person would choose in Exercise 7 if his utility function was $U(x)=x$ for $x\ge 0$.

10. Consider four gambles X_1, X_2, X_3, and X_4, for which the probability distributions of the gains are as follows:

$$\Pr(X_1=0)=0.2,\ \Pr(X_1=1)=0.5,\ \Pr(X_1=2)=0.3;$$
$$\Pr(X_2=0)=0.4,\ \Pr(X_2=1)=0.2,\ \Pr(X_2=2)=0.4;$$
$$\Pr(X_3=0)=0.3,\ \Pr(X_3=1)=0.3,\ \Pr(X_3=2)=0.4;$$
$$\Pr(X_4=0)=\Pr(X_4=2)=0.5.$$

Suppose that a person's utility function is such that she prefers X_1 to X_2. If the person were forced to accept either X_3 or X_4, which one would she choose?

11. Suppose that a person has a given fortune $A>0$ and can bet any amount b of this fortune in a certain game $(0\le b\le A)$. If he wins the bet, then his fortune becomes $A+b$; if he loses the bet, then his fortune becomes $A-b$. In general, let X denote his fortune after he has won or lost. Assume that the probability of his winning is p $(0<p<1)$ and the probability of his losing is $1-p$. Assume also that his utility function, as a function of his final fortune x, is $U(x)=\log x$ for $x>0$. If the person wishes to bet an amount b for which the expected utility of his fortune $E[U(X)]$ will be a maximum, what amount b should he bet?

12. Determine the amount b that the person should bet in Exercise 11 if his utility function is $U(x)=x^{1/2}$ for $x\ge 0$.

13. Determine the amount b that the person should bet in Exercise 11 if his utility function is $U(x)=x$ for $x\ge 0$.

14. Determine the amount b that the person should bet in Exercise 11 if his utility function is $U(x)=x^2$ for $x\ge 0$.

15. Suppose that a person has a lottery ticket from which she will win X dollars, where X has the uniform distribution on the interval [0, 4]. Suppose also that the person's utility function is $U(x)=x^{\alpha}$ for $x\ge 0$, where α is a given positive constant. For how many dollars x_0 would the person be willing to sell this lottery ticket?

16. Let Y be a random variable that we would like to predict. Suppose that we must choose a single number d as the prediction and that we will lose $(Y-d)^2$ dollars. Suppose that our utility for dollars is a square root function:

$$U(x)=\begin{cases}\sqrt{x} & \text{if } x\ge 0,\\ -\sqrt{-x} & \text{if } x<0.\end{cases}$$

Prove that the value of d that maximizes expected utility is a median of the distribution of Y.

17. Reconsider the conditions of Example 4.8.9. This time, suppose that $p_0=1/2$ and

$$U(x)=\begin{cases}x^{0.9} & \text{if } x\ge 0,\\ x & \text{if } x<0.\end{cases}$$

Suppose also that P has p.d.f. $f(p)=56p^6(1-p)$ for $0<p<1$. Decide whether or not it is better to abandon the treatment.

4.9 Supplementary Exercises

1. Suppose that the random variable X has a continuous distribution with c.d.f. $F(x)$ and p.d.f. f. Suppose also that $E(X)$ exists. Prove that

$$\lim_{x\to\infty} x[1 - F(x)] = 0.$$

Hint: Use the fact that if $E(X)$ exists, then

$$E(X) = \lim_{u\to\infty} \int_{-\infty}^{u} xf(x)\,dx.$$

2. Suppose that the random variable X has a continuous distribution with c.d.f. $F(x)$. Suppose also that $\Pr(X \ge 0) = 1$ and that $E(X)$ exists. Show that

$$E(X) = \int_0^\infty [1 - F(x)]\,dx.$$

Hint: You may use the result proven in Exercise 1.

3. Consider again the conditions of Exercise 2, but suppose now that X has a discrete distribution with c.d.f. $F(x)$, rather than a continuous distribution. Show that the conclusion of Exercise 2 still holds.

4. Suppose that X, Y, and Z are nonnegative random variables such that $\Pr(X + Y + Z \le 1.3) = 1$. Show that X, Y, and Z cannot possibly have a joint distribution under which each of their marginal distributions is the uniform distribution on the interval [0, 1].

5. Suppose that the random variable X has mean μ and variance σ^2, and that $Y = aX + b$. Determine the values of a and b for which $E(Y) = 0$ and $\text{Var}(Y) = 1$.

6. Determine the expectation of the range of a random sample of size n from the uniform distribution on the interval [0, 1].

7. Suppose that an automobile dealer pays an amount X (in thousands of dollars) for a used car and then sells it for an amount Y. Suppose that the random variables X and Y have the following joint p.d.f.:

$$f(x, y) = \begin{cases} \frac{1}{36}x & \text{for } 0 < x < y < 6, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the dealer's expected gain from the sale.

8. Suppose that $X_1, \dots, X_n$ form a random sample of size n from a continuous distribution with the following p.d.f.:

$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y_n = \max\{X_1, \dots, X_n\}$. Evaluate $E(Y_n)$.

9. If m is a median of the distribution of X, and if $Y = r(X)$ is either a nondecreasing or a nonincreasing function of X, show that $r(m)$ is a median of the distribution of Y.

10. Suppose that $X_1, \dots, X_n$ are i.i.d. random variables, each of which has a continuous distribution with median m. Let $Y_n = \max\{X_1, \dots, X_n\}$. Determine the value of $\Pr(Y_n > m)$.

11. Suppose that you are going to sell cola at a football game and must decide in advance how much to order. Suppose that the demand for cola at the game, in liters, has a continuous distribution with p.d.f. $f(x)$. Suppose that you make a profit of g cents on each liter that you sell at the game and suffer a loss of c cents on each liter that you order but do not sell. What is the optimal amount of cola for you to order so as to maximize your expected net gain?

12. Suppose that the number of hours X for which a machine will operate before it fails has a continuous distribution with p.d.f. $f(x)$. Suppose that at the time at which the machine begins operating you must decide when you will return to inspect it. If you return before the machine has failed, you incur a cost of b dollars for having wasted an inspection. If you return after the machine has failed, you incur a cost of c dollars per hour for the length of time during which the machine was not operating after its failure. What is the optimal number of hours to wait before you return for inspection in order to minimize your expected cost?

13. Suppose that X and Y are random variables for which $E(X) = 3$, $E(Y) = 1$, $\text{Var}(X) = 4$, and $\text{Var}(Y) = 9$. Let $Z = 5X - Y + 15$. Find $E(Z)$ and $\text{Var}(Z)$ under each of the following conditions: **(a)** X and Y are independent; **(b)** X and Y are uncorrelated; **(c)** the correlation of X and Y is 0.25.

14. Suppose that $X_0, X_1, \dots, X_n$ are independent random variables, each having the same variance σ^2. Let $Y_j = X_j - X_{j-1}$ for $j = 1, \dots, n$, and let $\overline{Y}_n = \frac{1}{n}\sum_{j=1}^n Y_j$. Determine the value of $\text{Var}(\overline{Y}_n)$.

15. Suppose that $X_1, \dots, X_n$ are random variables for which $\text{Var}(X_i)$ has the same value σ^2 for $i = 1, \dots, n$ and $\rho(X_i, X_j)$ has the same value ρ for every pair of values i and j such that $i \ne j$. Prove that $\rho \ge -\frac{1}{n-1}$.

16. Suppose that the joint distribution of X and Y is the uniform distribution over a rectangle with sides parallel to the coordinate axes in the xy-plane. Determine the correlation of X and Y.

17. Suppose that n letters are put at random into n envelopes, as in the matching problem described in Sec. 1.10. Determine the variance of the number of letters that are placed in the correct envelopes.

18. Suppose that the random variable X has mean μ and variance σ^2. Show that the third central moment of X can be expressed as $E(X^3) - 3\mu\sigma^2 - \mu^3$.

19. Suppose that X is a random variable with m.g.f. $\psi(t)$, mean μ, and variance σ^2; and let $c(t) = \log[\psi(t)]$. Prove that $c'(0) = \mu$ and $c''(0) = \sigma^2$.

20. Suppose that X and Y have a joint distribution with means μ_X and μ_Y, standard deviations σ_X and σ_Y, and correlation ρ. Show that if $E(Y|X)$ is a linear function of X, then

$$E(Y|X) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X).$$

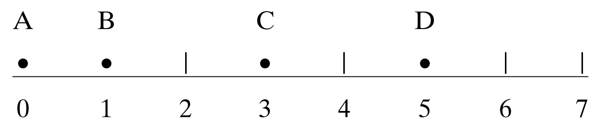
21. Suppose that X and Y are random variables such that $E(Y|X) = 7 - (1/4)X$ and $E(X|Y) = 10 - Y$. Determine the correlation of X and Y.

22. Suppose that a stick having a length of 3 feet is broken into two pieces, and that the point at which the stick is broken is chosen in accordance with the p.d.f. $f(x)$. What is the correlation between the length of the longer piece and the length of the shorter piece?

23. Suppose that X and Y have a joint distribution with correlation $\rho > 1/2$ and that $\text{Var}(X) = \text{Var}(Y) = 1$. Show that $b = -\dfrac{1}{2\rho}$ is the unique value of b such that the correlation of X and $X + bY$ is also ρ.

24. Suppose that four apartment buildings A, B, C, and D are located along a highway at the points 0, 1, 3, and 5, as shown in the following figure. Suppose also that 10 percent of the employees of a certain company live in building A, 20 percent live in B, 30 percent live in C, and 40 percent live in D.

- **a.** Where should the company build its new office in order to minimize the total distance that its employees must travel?
- **b.** Where should the company build its new office in order to minimize the sum of the squared distances that its employees must travel?



25. Suppose that X and Y have the following joint p.d.f.:

$$f(x, y) = \begin{cases} 8xy & \text{for } 0 < y < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose also that the observed value of X is 0.2.

- **a.** What predicted value of Y has the smallest M.S.E.?
- **b.** What predicted value of Y has the smallest M.A.E.?

26. For all random variables X, Y, and Z, let $\text{Cov}(X, Y|z)$ denote the covariance of X and Y in their conditional joint distribution given $Z = z$. Prove that

$$\text{Cov}(X, Y) = E[\text{Cov}(X, Y|Z)] + \text{Cov}[E(X|Z), E(Y|Z)].$$

27. Consider the box of red and blue balls in Examples 4.2.4 and 4.2.5. Suppose that we sample $n > 1$ balls with replacement, and let X be the number of red balls in the sample. Then we sample n balls without replacement, and we let Y be the number of red balls in the sample. Prove that $\Pr(X = n) > \Pr(Y = n)$.

28. Suppose that a person's utility function is $U(x) = x^2$ for $x \geq 0$. Show that the person will always prefer to take a gamble in which she will receive a random gain of X dollars rather than receive the amount $E(X)$ with certainty, where $\Pr(X \geq 0) = 1$ and $E(X) < \infty$.

29. A person is given m dollars, which he must allocate between an event A and its complement A^c. Suppose that he allocates a dollars to A and $m - a$ dollars to A^c. The person's gain is then determined as follows: If A occurs, his gain is $g_1 a$; if A^c occurs, his gain is $g_2(m - a)$. Here, g_1 and g_2 are given positive constants. Suppose also that $\Pr(A) = p$ and the person's utility function is $U(x) = \log x$ for $x > 0$. Determine the amount a that will maximize the person's expected utility, and show that this amount does not depend on the values of g_1 and g_2.

This page intentionally left blank

Chapter 5

Special Distributions

5.1 Introduction

In this chapter, we shall define and discuss several special families of distributions that are widely used in applications of probability and statistics. The distributions that will be presented here include discrete and continuous distributions of univariate, bivariate, and multivariate types. The discrete univariate distributions are the families of Bernoulli, binomial, hypergeometric, Poisson, negative binomial, and geometric distributions. The continuous univariate distributions are the families of normal, lognormal, gamma, exponential, and beta distributions. Other continuous univariate distributions (introduced in exercises and examples) are the families of Weibull and Pareto distributions. Also discussed is the multinomial family of multivariate discrete distributions, and the bivariate normal family of bivariate continuous distributions.

We shall briefly describe how each of these families of distributions arise in applied problems and show why each might be an appropriate probability model for some experiment. For each family, we shall present the form of the p.f. or the p.d.f. and discuss some of the basic properties of the distributions in the family.

The list of distributions presented in this chapter, or in this entire text for that matter, is not intended to be exhaustive. These distributions are known to be useful in a wide variety of applied problems. In many real-world problems, however, one will need to consider other distributions not mentioned here. The tools that we develop for use with these distributions can be generalized for use with other distributions. Our purpose in providing in-depth presentations of the most popular distributions here is to give the reader a feel for how to use probablity to model the variation and uncertainty in applied problems as well as some of the tools that get used during probability modeling.

5.2 The Bernoulli and Binomial Distributions

The simplest type of experiment has only two possible outcomes, call them 0 and 1. If X equals the outcome from such an experiment, then X has the simplest type of nondegenerate distribution, which is a member of the family of Bernoulli distributions. If n independent random variables $X_1, \dots, X_n$ *all have the same*

Bernoulli distribution, then their sum is equal to the number of the X_i's that equal 1, and the distribution of the sum is a member of the binomial family.

The Bernoulli Distributions

Example 5.2.1 A Clinical Trial. The treatment given to a particular patient in a clinical trial can either succeed or fail. Let $X = 0$ if the treatment fails, and let $X = 1$ if the treatment succeeds. All that is needed to specify the distribution of X is the value $p = \Pr(X = 1)$ (or, equivalently, $1 - p = \Pr(X = 0)$). Each different p corresponds to a different distribution for X. The collection of all such distributions corresponding to all $0 \le p \le 1$ form the family of Bernoulli distributions. ◀

An experiment of a particularly simple type is one in which there are only two possible outcomes, such as head or tail, success or failure, defective or nondefective, patient recovers or does not recover. It is convenient to designate the two possible outcomes of such an experiment as 0 and 1, as in Example 5.2.1. The following recap of Definition 3.1.5 can then be applied to every experiment of this type.

Definition 5.2.1 Bernoulli Distribution. A random variable X has the *Bernoulli distribution with parameter* p $(0 \le p \le 1)$ if X can take only the values 0 and 1 and the probabilities are

$$\Pr(X = 1) = p \quad \text{and} \quad \Pr(X = 0) = 1 - p. \tag{5.2.1}$$

The p.f. of X can be written as follows:

$$f(x|p) = \begin{cases} p^x(1-p)^{1-x} & \text{for } x = 0, 1, \\ 0 & \text{otherwise.} \end{cases} \tag{5.2.2}$$

To verify that this p.f. $f(x|p)$ actually does represent the Bernoulli distribution specified by the probabilities (5.2.1), it is simply necessary to note that $f(1|p) = p$ and $f(0|p) = 1 - p$.

If X has the Bernoulli distribution with parameter p, then X^2 and X are the same random variable. It follows that

$$E(X) = 1 \cdot p + 0 \cdot (1 - p) = p,$$
$$E(X^2) = E(X) = p,$$

and

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = p(1 - p).$$

Furthermore, the m.g.f. of X is

$$\psi(t) = E(e^{tX}) = pe^t + (1 - p) \quad \text{for } -\infty < t < \infty.$$

Definition 5.2.2 Bernoulli Trials/Process. If the random variables in a finite or infinite sequence $X_1, X_2, \ldots$ are i.i.d., and if each random variable X_i has the Bernoulli distribution with parameter p, then it is said that $X_1, X_2, \ldots$ are *Bernoulli trials with parameter* p. An infinite sequence of Bernoulli trials is also called a *Bernoulli process*.

Example 5.2.2 Tossing a Coin. Suppose that a fair coin is tossed repeatedly. Let $X_i = 1$ if a head is obtained on the ith toss, and let $X_i = 0$ if a tail is obtained $(i = 1, 2, \ldots)$. Then the random variables $X_1, X_2, \ldots$ are Bernoulli trials with parameter $p = 1/2$. ◀

Example 5.2.3 Defective Parts. Suppose that 10 percent of the items produced by a certain machine are defective and the parts are independent of each other. We will sample n items at random and inspect them. Let $X_i = 1$ if the ith item is defective, and let $X_i = 0$ if it is nondefective ($i = 1, \dots, n$). Then the variables $X_1, \dots, X_n$ form n Bernoulli trials with parameter $p = 1/10$. ◀

Example 5.2.4 Clinical Trials. In the many clinical trial examples in earlier chapters (Example 4.7.8, for instance), the random variables $X_1, X_2, \dots,$ indicating whether each patient is a success, were conditionally Bernoulli trials with parameter p given $P = p$, where P is the unknown proportion of patients in a very large population who recover. ◀

The Binomial Distributions

Example 5.2.5 Defective Parts. In Example 5.2.3, let $X = X_1 + \cdots + X_{10}$, which equals the number of defective parts among the 10 sampled parts. What is the distribution of X? ◀

As derived after Example 3.1.9, the distribution of X in Example 5.2.5 is the binomial distribution with parameters 10 and 1/10. We repeat the general definition of binomial distributions here.

Definition 5.2.3 Binomial Distribution. A random variable X has the *binomial distribution with parameters n and p* if X has a discrete distribution for which the p.f. is as follows:

$$f(x|n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, 2, \dots, n, \\ 0 & \text{otherwise.} \end{cases} \tag{5.2.3}$$

In this distribution, n must be a positive integer, and p must lie in the interval $0 \le p \le 1$.

Probabilities for various binomial distributions can be obtained from the table given at the end of this book and from many statistical software programs.

The binomial distributions are of fundamental importance in probability and statistics because of the following result, which was derived in Sec. 3.1 and which we restate here in the terminology of this chapter.

Theorem 5.2.1 If the random variables $X_1, \dots, X_n$ form n Bernoulli trials with parameter p, and if $X = X_1 + \cdots + X_n$, then X has the binomial distribution with parameters n and p. ■

When X is represented as the sum of n Bernoulli trials as in Theorem 5.2.1, the values of the mean, variance, and m.g.f. of X can be derived very easily. These values, which were already obtained in Example 4.2.5 and on pages 231 and 238, are

$$E(X) = \sum_{i=1}^{n} E(X_i) = np,$$

$$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i) = np(1-p),$$

and

$$\psi(t) = E(e^{tX}) = \prod_{i=1}^{n} E(e^{tX_i}) = (pe^t + 1 - p)^n. \tag{5.2.4}$$

The reader can use the m.g.f. in Eq. (5.2.4) to establish the following simple extension of Theorem 4.4.6.

Theorem 5.2.2 If $X_1, \dots, X_k$ are independent random variables, and if X_i has the binomial distribution with parameters n_i and p $(i = 1, \dots, k)$, then the sum $X_1 + \cdots + X_k$ has the binomial distribution with parameters $n = n_1 + \cdots + n_k$ and p. ■

Theorem 5.2.2 also follows easily if we represent each X_i as the sum of n_i Bernoulli trials with parameter p. If $n = n_1 + \cdots + n_k$, and if all n trials are independent, then the sum $X_1 + \cdots + X_k$ will simply be the sum of n Bernoulli trials with parameter p. Hence, this sum must have the binomial distribution with parameters n and p.

Example 5.2.6 Castaneda v. Partida. Courts have used the binomial distributions to calculate probabilities of jury compositions from populations with known racial and ethnic compositions. In the case of *Castaneda v. Partida*, 430 U.S. 482 (1977), a local population was 79.1 percent Mexican American. During a 2.5-year period, there were 220 persons called to serve on grand juries, but only 100 were Mexican Americans. The claim was made that this was evidence of discrimination against Mexican Americans in the grand jury selection process. The court did a calculation under the assumption that grand jurors were drawn at random and independently from the population each with probability 0.791 of being Mexican American. Since the claim was that 100 was too small a number of Mexican Americans, the court calculated the probability that a binomial random variable X with parameters 220 and 0.791 would be 100 or less. The probability is very small (less than 10^{-25}). Is this evidence of discrimination against Mexican Americans? The small probability was calculated under the assumption that X had the binomial distribution with parameters 220 and 0.791, which means that the court was assuming that there was no discrimination against Mexican Americans when performing the calculation. In other words, the small probability is the conditional probability of observing $X \le 100$ given that there is no discrimination. What should be more interesting to the court is the reverse conditional probability, namely, the probability that there is no discrimination given that $X = 100$ (or given $X \le 100$). This sounds like a case for Bayes' theorem. After we introduce the beta distributions in Sec. 5.8, we shall show how to use Bayes' theorem to calculate this probability (Examples 5.8.3 and 5.8.4). ◀

Note: Bernoulli and Binomial Distributions. Every random variable that takes only the two values 0 and 1 must have a Bernoulli distribution. However, not every sum of Bernoulli random variables has a binomial distribution. There are two conditions needed to apply Theorem 5.2.1. The Bernoulli random variables must be mutually independent, and they must all have the same parameter. If either of these conditions fails, the distribution of the sum will not be a binomial distribution. When the court did a binomial calculation in Example 5.2.6, it was defining "no discrimination" to mean that jurors were selected independently and with the same probability 0.791 of being Mexican American. If the court had defined "no discrimination" some other way, they would have needed to do a different, presumably more complicated, probability calculation.

We conclude this section with an example that shows how Bernoulli and binomial calculations can improve efficiency when data collection is costly.

Example 5.2.7 Group Testing. Military and other large organizations are often faced with the need to test large numbers of members for rare diseases. Suppose that each test requires

a small amount of blood, and it is guaranteed to detect the disease if it is anywhere in the blood. Suppose that 1000 people need to be tested for a disease that affects 1/5 of 1 percent of all people. Let $X_j = 1$ if person j has the disease and $X_j = 0$ if not, for $j = 1, \ldots, 1000$. We model the X_j as i.i.d. Bernoulli random variables with parameter 0.002 for $j = 1, \ldots, 1000$. The most naïve approach would be to perform 1000 tests to see who has the disease. But if the tests are costly, there may be a more economical way to test. For example, one could divide the 1000 people into 10 groups of size 100 each. For each group, take a portion of the blood sample from each of the 100 people in the group and combine them into one sample. Then test each of the 10 combined samples. If none of the 10 combined samples has the disease, then nobody has the disease, and we needed only 10 tests instead of 1000. If only one of the combined samples has the disease, then we can test those 100 people separately, and we needed only 110 tests.

In general, let $Z_{1,i}$ be the number of people in group i who have the disease for $i = 1, \ldots, 10$. Then each $Z_{1,i}$ has the binomial distribution with parameters 100 and 0.002. Let $Y_{1,i} = 1$ if $Z_{1,i} > 0$ and $Y_{1,i} = 0$ if $Z_{1,i} = 0$. Then each $Y_{1,i}$ has the Bernoulli distribution with parameter

$$\Pr(Z_{1,i} > 0) = 1 - \Pr(Z_{1,i} = 0) = 1 - 0.998^{100} = 0.181,$$

and they are independent. Then $Y_1 = \sum_{i=1}^{10} Y_{1,i}$ is the number of groups whose members we have to test individually. Also, Y_1 has the binomial distribution with parameters 10 and 0.181. The number of people that we need to test individually is $100Y_1$. The mean of $100Y_1$ is $100 \times 10 \times 0.181 = 181$. So, the expected total number of tests is $10 + 181 = 191$, rather than 1000. One can compute the entire distribution of the total number of tests, $100Y_1 + 10$. The maximum number of tests needed by this group testing procedure is 1010, which would be the case if all 10 groups had at least one person with the disease, but this has probability 3.84×10^{-8}. In all other cases, group testing requires fewer than 1000 tests.

There are multiple-stage versions of group testing in which each of the groups that tests positive is split further into subgroups which are each tested together. If each of those subgroups is sufficiently large, they can be further subdivided into smaller sub-subgroups, etc. Finally, only the final-stage subgroups that have a positive result are tested individually. This can further reduce the expected number of tests. For example, consider the following two-stage version of the procedure described earlier. We could divide each of the 10 groups of 100 people into 10 subgroups of 10 people each. Following the above notation, let $Z_{2,i,k}$ be the number of people in subgroup k of group i who have the disease, for $i = 1, \ldots, 10$ and $k = 1, \ldots, 10$. Then each $Z_{2,i,k}$ has the binomial distribution with parameters 10 and 0.002. Let $Y_{2,i,k} = 1$ if $Z_{2,i,k} > 0$ and $Y_{2,i,k} = 0$ otherwise. Notice that $Y_{2,i,k} = 0$ for $k = 1, \ldots, 10$ for every i such that $Y_{1,i} = 0$. So, we only need to test individuals in those subgroups such that $Y_{2,i,k} = 1$. Each $Y_{2,i,k}$ has the Bernoulli distribution with parameter

$$\Pr(Z_{2,i,k} > 0) = 1 - \Pr(Z_{2,i,k} = 0) = 1 - 0.998^{10} = 0.0198,$$

and they are independent. Then $Y_2 = \sum_{i=1}^{10} \sum_{j=1}^{10} Y_{2,i,k}$ is the number of groups whose members we have to test individually. Also, Y_2 has the binomial distribution with parameters 100 and 0.0198. The number of people that we need to test individually is $10Y_2$. The mean of $10Y_2$ is $10 \times 100 \times 0.0198 = 19.82$. The number of subgroups that we need to test in the second stage is Y_1, whose mean is 1.81. So, the expected total number of tests is $10 + 1.81 + 19.82 = 31.63$, which is even smaller than the 191 for the one-stage procedure described earlier. ◀

Summary

A random variable X has the Bernoulli distribution with parameter p if the p.f. of X is $f(x|p) = p^x(1-p)^{1-x}$ for $x = 0, 1$ and 0 otherwise. If $X_1, \dots, X_n$ are i.i.d. random variables all having the Bernoulli distribution with parameter p, then we refer to $X_1, \dots, X_n$ as Bernoulli trials, and $X = \sum_{i=1}^n X_i$ has the binomial distribution with parameters n and p. Also, X is the number of successes in the n Bernoulli trials, where success on trial i corresponds to $X_i = 1$ and failure corresponds to $X_i = 0$.

Exercises

1. Suppose that X is a random variable such that $E(X^k) = 1/3$ for $k = 1, 2, \dots$. Assuming that there cannot be more than one distribution with this same sequence of moments (see Exercise 14), determine the distribution of X.

2. Suppose that a random variable X can take only the two values a and b with the following probabilities:

$$\Pr(X = a) = p \quad \text{and} \quad \Pr(X = b) = 1 - p.$$

Express the p.f. of X in a form similar to that given in Eq. (5.2.2).

3. Suppose that a fair coin (probability of heads equals 1/2) is tossed independently 10 times. Use the table of the binomial distribution given at the end of this book to find the probability that strictly more heads are obtained than tails.

4. Suppose that the probability that a certain experiment will be successful is 0.4, and let X denote the number of successes that are obtained in 15 independent performances of the experiment. Use the table of the binomial distribution given at the end of this book to determine the value of $\Pr(6 \le X \le 9)$.

5. A coin for which the probability of heads is 0.6 is tossed nine times. Use the table of the binomial distribution given at the end of this book to find the probability of obtaining an even number of heads.

6. Three men A, B, and C shoot at a target. Suppose that A shoots three times and the probability that he will hit the target on any given shot is 1/8, B shoots five times and the probability that he will hit the target on any given shot is 1/4, and C shoots twice and the probability that he will hit the target on any given shot is 1/2. What is the expected number of times that the target will be hit?

7. Under the conditions of Exercise 6, assume also that all shots at the target are independent. What is the variance of the number of times that the target will be hit?

8. A certain electronic system contains 10 components. Suppose that the probability that each individual component will fail is 0.2 and that the components fail independently of each other. Given that at least one of the components has failed, what is the probability that at least two of the components have failed?

9. Suppose that the random variables $X_1, \dots, X_n$ form n Bernoulli trials with parameter p. Determine the conditional probability that $X_1 = 1$, given that

$$\sum_{i=1}^{n} X_i = k \quad (k = 1, \dots, n).$$

10. The probability that each specific child in a given family will inherit a certain disease is p. If it is known that at least one child in a family of n children has inherited the disease, what is the expected number of children in the family who have inherited the disease?

11. For $0 \le p \le 1$, and $n = 2, 3, \dots$, determine the value of

$$\sum_{x=2}^{n} x(x-1)\binom{n}{x} p^x (1-p)^{n-x}.$$

12. If a random variable X has a discrete distribution for which the p.f. is $f(x)$, then the value of x for which $f(x)$ is maximum is called the *mode* of the distribution. If this same maximum $f(x)$ is attained at more than one value of x, then all such values of x are called *modes* of the distribution. Find the mode or modes of the binomial distribution with parameters n and p. *Hint:* Study the ratio $f(x+1|n, p)/f(x|n, p)$.

13. In a clinical trial with two treatment groups, the probability of success in one treatment group is 0.5, and the probability of success in the other is 0.6. Suppose that there are five patients in each group. Assume that the outcomes of all patients are independent. Calculate the probability that the first group will have at least as many successes as the second group.

14. In Exercise 1, we assumed that there could be at most one distribution with moments $E(X^k) = 1/3$ for $k = 1, 2, \dots$. In this exercise, we shall prove that there can be only one such distribution. Prove the following

facts and show that they imply that at most one distribution has the given moments.

a. $\Pr(|X| \le 1) = 1$. (If not, show that $\lim_{k\to\infty} E(X^{2k}) = \infty$.)

b. $\Pr(X^2 \in \{0, 1\}) = 1$. (If not, prove that $E(X^4) < E(X^2)$.)

c. $\Pr(X = -1) = 0$. (If not, prove that $E(X) < E(X^2)$.)

15. In Example 5.2.7, suppose that we use the two-stage version described at the end of the example. What is the maximum number of tests that could possibly be needed by this version? What is the probability that the maximum number of tests would be required?

16. For the 1000 people in Example 5.2.7, suppose that we use the following three-stage group testing procedure. First, divide the 1000 people into five groups of size 200 each. For each group that tests positive, further divide it into five subgroups of size 40 each. For each subgroup that tests positive, further divide it into five sub-subgroups of size 8 each. For each sub-subgroup that tests positive, test all eight people. Find the expected number and maximum number of tests.

5.3 The Hypergeometric Distributions

In this section, we consider dependent Bernoulli random variables. A common source of dependent Bernoulli random variables is sampling without replacement from a finite population. Suppose that a finite population consists of a known number of successes and failures. If we sample a fixed number of units from that population, the number of successes in our sample will have a distribution that is a member of the family of hypergeometric distributions.

Definition and Examples

Example 5.3.1

Sampling without Replacement. Suppose that a box contains A red balls and B blue balls. Suppose also that $n \ge 0$ balls are selected at random from the box without replacement, and let X denote the number of red balls that are obtained. Clearly, we must have $n \le A + B$ or we would run out of balls. Also, if $n = 0$, then $X = 0$ because there are no balls, red or blue, drawn. For cases with $n \ge 1$, we can let $X_i = 1$ if the ith ball drawn is red and $X_i = 0$ if not. Then each X_i has a Bernoulli distribution, but $X_1, \dots, X_n$ are not independent in general. To see this, assume that both $A > 0$ and $B > 0$ as well as $n \ge 2$. We will now show that $\Pr(X_2 = 1|X_1 = 0) \ne \Pr(X_2 = 1|X_1 = 1)$. If $X_1 = 1$, then when the second ball is drawn there are only $A - 1$ red balls remaining out of a total of $A + B - 1$ available balls. Hence, $\Pr(X_2 = 1|X_1 = 1) = (A - 1)/(A + B - 1)$. By the same reasoning,

$$\Pr(X_2 = 1|X_1 = 0) = \frac{A}{A + B - 1} > \frac{A - 1}{A + B - 1}.$$

Hence, X_2 is not independent of X_1, and we should not expect X to have a binomial distribution. ◀

The problem described in Example 5.3.1 is a template for all cases of sampling without replacement from a finite population with only two types of objects. Anything that we learn about the random variable X in Example 5.3.1 will apply to every case of sampling without replacement from finite populations with only two types of objects. First, we derive the distribution of X.

Theorem 5.3.1 Probability Function. The distribution of X in Example 5.3.1 has the p.f.

$$f(x|A, B, n) = \frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}}, \tag{5.3.1}$$

for

$$\max\{0, n - B\} \le x \le \min\{n, A\}, \tag{5.3.2}$$

and $f(x|A, B, n) = 0$ otherwise.

Proof Clearly, the value of X can neither exceed n nor exceed A. Therefore, it must be true that $X \le \min\{n, A\}$. Similarly, because the number of blue balls $n - X$ that are drawn cannot exceed B, the value of X must be at least $n - B$. Because the value of X cannot be less than 0, it must be true that $X \ge \max\{0, n - B\}$. Hence, the value of X must be an integer in the interval in (5.3.2).

We shall now find the p.f. of X using combinatorial arguments from Sec. 1.8. The degenerate cases, those with A, B, and/or n equal to 0, are easy to prove because $\binom{k}{0} = 1$ for all nonnegative k, including $k = 0$. For the cases in which all of A, B, and n are strictly positive, there are $\binom{A+B}{n}$ ways to choose n balls out of the $A + B$ available balls, and all of these choices are equally likely. For each integer x in the interval (5.3.2), there are $\binom{A}{x}$ ways to choose x red balls, and for each such choice there are $\binom{B}{n-x}$ ways to choose $n - x$ blue balls. Hence, the probability of obtaining exactly x red balls out of n is given by Eq. (5.3.1). Furthermore, $f(x|A, B, n)$ must be 0 for all other values of x, because all other values are impossible. ■

Definition 5.3.1 Hypergeometric Distribution. Let A, B, and n be nonnegative integers with $n \le A + B$. If a random variable X has a discrete distribution with p.f. as in Eqs. (5.3.1) and (5.3.2), then it is said that X has the *hypergeometric distribution with parameters A, B, and n*.

Example 5.3.2 Sampling without Replacement from an Observed Data Set. Consider the patients in the clinical trial whose results are tabulated in Table 2.1. We might need to reexamine a subset of the patients in the placebo group. Suppose that we need to sample 11 distinct patients from the 34 patients in that group. What is the distribution of the number of successes (no relapse) that we obtain in the subsample? Let X stand for the number of successes in the subsample. Table 2.1 indicates that there are 10 successes and 24 failures in the placebo group. According to the definition of the hypergeometric distribution, X has the hypergeometric distribution with parameters $A = 10$, $B = 24$, and $n = 11$. In particular, the possible values of X are the integers from 0 to 10. Even though we sample 11 patients, we cannot observe 11 successes, since only 10 successes are available. ◄

The Mean and Variance for a Hypergeometric Distribution

Theorem 5.3.2 Mean and Variance. Let X have a hypergeometric distribution with strictly positive parameters A, B, and n. Then

$$E(X) = \frac{nA}{A+B}, \tag{5.3.3}$$

$$\text{Var}(X) = \frac{nAB}{(A+B)^2} \cdot \frac{A+B-n}{A+B-1}. \tag{5.3.4}$$

Proof Assume that X is as defined in Example 5.3.1, the number of red balls drawn when n balls are selected at random without replacement from a box containing A red balls and B blue balls. For $i = 1, \ldots, n$, let $X_i = 1$ if the ith ball that is selected is red, and let $X_i = 0$ if the ith ball is blue. As explained in Example 4.2.4, we can imagine that the n balls are selected from the box by first arranging all the balls in the box in some random order and then selecting the first n balls from this arrangement. It can be seen from this interpretation that, for $i = 1, \ldots, n$,

$$\Pr(X_i = 1) = \frac{A}{A+B} \quad \text{and} \quad \Pr(X_i = 0) = \frac{B}{A+B}.$$

Therefore, for $i = 1, \ldots, n$,

$$E(X_i) = \frac{A}{A+B} \quad \text{and} \quad \text{Var}(X_i) = \frac{AB}{(A+B)^2}. \tag{5.3.5}$$

Since $X = X_1 + \cdots + X_n$, the mean of X is the sum of the means of the X_i's, namely, Eq. (5.3.3).

Next, use Theorem 4.6.7 to write

$$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{i<j}\sum \text{Cov}(X_i, X_j). \tag{5.3.6}$$

Because of the symmetry among the random variables $X_1, \ldots, X_n$, every term $\text{Cov}(X_i, X_j)$ in the final summation in Eq. (5.3.6) will have the same value as $\text{Cov}(X_1, X_2)$. Since there are $\binom{n}{2}$ terms in this summation, it follows from Eqs. (5.3.5) and (5.3.6) that

$$\text{Var}(X) = \frac{nAB}{(A+B)^2} + n(n-1)\,\text{Cov}(X_1, X_2). \tag{5.3.7}$$

We could compute $\text{Cov}(X_1, X_2)$ directly, but it is simpler to argue as follows. If $n = A + B$, then $\Pr(X = A) = 1$ because *all* the balls in the box will be selected without replacement. Thus, for $n = A + B$, X is a constant random variable and $\text{Var}(X) = 0$. Setting Eq. (5.3.7) to 0 and solving for $\text{Cov}(X_1, X_2)$ gives

$$\text{Cov}(X_1, X_2) = -\frac{AB}{(A+B)^2(A+B-1)}.$$

Plugging this value back into Eq. (5.3.7) gives Eq. (5.3.4). ■

Comparison of Sampling Methods

If we had sampled *with* replacement in Example 5.3.1, the number of red balls would have the binomial distribution with parameters n and $A/(A+B)$. In that case, the mean number of red balls would still be $nA/(A+B)$, but the variance would be different. To see how the variances from sampling with and without replacement are related, let $T = A + B$ denote the total number of balls in the box, and let $p = A/T$ denote the proportion of red balls in the box. Then Eq. (5.3.4) can be rewritten as follows:

$$\text{Var}(X) = np(1-p)\frac{T-n}{T-1}. \tag{5.3.8}$$

The variance $np(1-p)$ of the binomial distribution is the variance of the number of red balls when sampling with replacement. The factor $\alpha = (T-n)/(T-1)$ in Eq. (5.3.8) therefore represents the reduction in $\text{Var}(X)$ caused by sampling without replacement from a finite population. This α is called the *finite population correction* in the theory of sampling from finite populations without replacement.

If $n = 1$, the value of this factor α is 1, because there is no distinction between sampling with replacement and sampling without replacement when only one ball is being selected. If $n = T$, then (as previously mentioned) $\alpha = 0$ and $\text{Var}(X) = 0$. For values of n between 1 and T, the value of α will be between 0 and 1.

For each fixed sample size n, it can be seen that $\alpha \to 1$ as $T \to \infty$. This limit reflects the fact that when the population size T is very large compared to the sample size n, there is very little difference between sampling with replacement and sampling without replacement. Theorem 5.3.4 expresses this idea more formally. The proof relies on the following result which gets used several times in this text.

Theorem 5.3.3 Let a_n and c_n be sequences of real numbers such that a_n converges to 0, and $c_n a_n^2$ converges to 0. Then

$$\lim_{n\to\infty} (1+a_n)^{c_n} e^{-a_n c_n} = 1.$$

In particular, if $a_n c_n$ converges to b, then $(1+a_n)^{c_n}$ converges to e^b. ■

The proof of Theorem 5.3.3 is left to the reader in Exercise 11.

Theorem 5.3.4 **Closeness of Binomial and Hypergeometric Distributions.** Let $0 < p < 1$, and let n be a positive integer. Let Y have the binomial distribution with parameters n and p. For each positive integer T, let A_T and B_T be integers such that $\lim_{T\to\infty} A_T = \infty$, $\lim_{T\to\infty} B_T = \infty$, and $\lim_{T\to\infty} A_T/(A_T + B_T) = p$. Let X_T have the hypergeometric distribution with parameters A_T, B_T, and n. For each fixed n and each $x = 0, \dots, n$,

$$\lim_{T\to\infty} \frac{\Pr(Y = x)}{\Pr(X_T = x)} = 1. \tag{5.3.9}$$

Proof Once A_T and B_T are both larger than n, the formula in (5.3.1) is $\Pr(X_T = x)$ for all $x = 0, \dots, n$. So, for large T, we have

$$\Pr(X_T = x) = \binom{n}{x} \frac{A_T! B_T! (A_T + B_T - n)!}{(A_T - x)!(B_T - n + x)!(A_T + B_T)!}.$$

Apply Stirling's formula (Theorem 1.7.5) to each of the six factorials in the second factor above. A little manipulation gives that

$$\lim_{T\to\infty} \frac{\binom{n}{x} A_T^{A_T+1/2} B_T^{B_T+1/2} (A_T + B_T - n)^{A_T+B_T-n+1/2}}{\Pr(X_T = x)(A_T - x)^{A_T-x+1/2}(B_T - n + x)^{B_T-n+x+1/2}(A_T + B_T)^{A_T+B_T+1/2}} \tag{5.3.10}$$

equals 1. Each of the following limits follows from Theorem 5.3.3:

$$\lim_{T\to\infty} \left(\frac{A_T}{A_T - x}\right)^{A_T-x+1/2} = e^x$$

$$\lim_{T\to\infty} \left(\frac{B_T}{B_T - n + x}\right)^{B_T-n+x+1/2} = e^{n-x}$$

$$\lim_{T\to\infty} \left(\frac{A_T + B_T - n}{A_T + B_T}\right)^{A_T+B_T-n+1/2} = e^{-n}.$$

Inserting these limits in (5.3.10) yields

$$\lim_{T\to\infty} \frac{\binom{n}{x} A_T^x B_T^{n-x}}{\Pr(X_T = x)(A_T + B_T)^n} = 1. \tag{5.3.11}$$

Since $A_T/(A_T + B_T)$ converges to p, we have

$$\lim_{T\to\infty} \frac{A_T^x B_T^{n-x}}{(A_T + B_T)^n} = p^x(1-p)^{n-x}. \tag{5.3.12}$$

Together, (5.3.11) and (5.3.12) imply that

$$\lim_{T\to\infty} \frac{\binom{n}{x} p^x (1-p)^{n-x}}{\Pr(X_T = x)} = 1.$$

The numerator of this last expression is $\Pr(Y = x)$; hence, (5.3.9) holds. ■

In words, Theorem 5.3.4 says that if the sample size n represents a negligible fraction of the total population $A + B$, then the hypergeometric distribution with parameters A, B, and n will be very nearly the same as the binomial distribution with parameters n and $p = A/(A + B)$.

Example 5.3.3

Population of Unknown Composition. The hypergeometric distribution can arise as a conditional distribution when sampling is done without replacement from a finite population of unknown composition. The simplest example would be to modify Example 5.3.1 so that we still know the value of $T = A + B$ but no longer know A and B. That is, we know how many balls are in the box, but we don't know how many are red or blue. This makes $P = A/T$, the proportion of red balls, unknown. Let $h(p)$ be the p.f. of P. Here P is a random variable whose possible values are $0, 1/T, \ldots, (T-1)/T, 1$. Conditional on $P = p$, we can behave as if we know that $A = pT$ and $B = (1-p)T$, and then the conditional distribution of X (the number of red balls in a sample of size n) is the hypergeometric distribution with parameters pT, $(1-p)T$, and n.

Suppose now that T is so large that the difference is essentially negligible between this hypergeometric distribution and the binomial distribution with parameters n and p. In this case, it is no longer necessary that we assume that T is known. This is the situation that we had in mind (in Examples 3.4.10 and 3.6.7, as well as their many variations and other examples) when we referred to P as the proportion of successes among all patients who might receive a treatment or the proportion of defectives among all parts produced by a machine. We think of T as essentially infinite so that conditional on the proportion A/T, which we call P, the individual draws become independent Bernoulli trials. If either A or T (or both) is unknown, it makes sense that $P = A/T$ will be unknown. In the augmented experiment described on page 61, in which P can be computed from the experimental outcome, we have that P is a random variable. ◀

Note: Essentially Infinite Populations. The case in which T is essentially infinite in Example 5.3.3 is the motivation for using the binomial distributions as models for numbers of successes in samples from very large finite populations. Look at Example 5.2.6, for instance. The number of Mexican Americans available to be sampled for grand jury duty is finite, but it is huge relative to the number (220) of grand jurors selected during the 2.5-year period. Technically, it is impossible that the individual grand jurors are selected independently, but the difference is too small for even the best defense attorney to make anything out of it. In the future, we will often model Bernoulli random variables as independent when we imagine selecting them

at random without replacement from a huge finite population. We shall be relying on Theorem 5.3.4 in these cases without explicitly saying so.

Extending the Definition of Binomial Coefficients

There is an extension of the definition of a binomial coefficient given in Sec. 1.8 that allows a simplification of the expression for the p.f. of the hypergeometric distribution. For all positive integers r and m, where $r \le m$, the binomial coefficient $\binom{m}{r}$ was defined to be

$$\binom{m}{r} = \frac{m!}{r!(m-r)!}. \tag{5.3.13}$$

It can be seen that the value of $\binom{m}{r}$ specified by Eq. (5.3.13) can also be written in the form

$$\binom{m}{r} = \frac{m(m-1)\cdots(m-r+1)}{r!}. \tag{5.3.14}$$

For every real number m that is not necessarily a positive integer and every positive integer r, the value of the right side of Eq. (5.3.14) is a well-defined number. Therefore, for every real number m and every positive integer r, we can extend the definition of the binomial coefficient $\binom{m}{r}$ by defining its value as that given by Eq. (5.3.14).

The value of the binomial coefficient $\binom{m}{r}$ can be obtained from this definition for all positive integers r and m. If $r \le m$, the value of $\binom{m}{r}$ is given by Eq.(5.3.13). If $r > m$, one of the factors in the numerator of (5.3.14) will be 0 and $\binom{m}{r} = 0$. Finally, for every real number m, we shall define the value of $\binom{m}{0}$ to be $\binom{m}{0} = 1$.

When this extended definition of a binomial coefficient is used, it can be seen that the value of $\binom{A}{x}\binom{B}{n-x}$ is 0 for every integer x such that either $x > A$ or $n - x > B$. Therefore, we can write the p.f. of the hypergeometric distribution with parameters A, B, and n as follows:

$$f(x|A, B, n) = \begin{cases} \dfrac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} & \text{for } x = 0, 1, \ldots, n, \\ 0 & \text{otherwise.} \end{cases} \tag{5.3.15}$$

It then follows from Eq. (5.3.14) that $f(x|A, B, n) > 0$ if and only if x is an integer in the interval (5.3.2).

Summary

We introduced the family of hypergeometric distributions. Suppose that n units are drawn at random without replacement from a finite population consisting of T units of which A are successes and $B = T - A$ are failures. Let X stand for the number of successes in the sample. Then the distribution of X is the hypergeometric distribution with parameters A, B, and n. We saw that the distinction between sampling from a finite population with and without replacement is negligible when the size of the population is huge relative to the size of the sample. We also generalized the binomial coefficient notation so that $\binom{m}{r}$ is defined for all real numbers m and all positive integers r.

Exercises

1. In Example 5.3.2, compute the probability that all 10 success patients appear in the subsample of size 11 from the Placebo group.

2. Suppose that a box contains five red balls and ten blue balls. If seven balls are selected at random without replacement, what is the probability that at least three red balls will be obtained?

3. Suppose that seven balls are selected at random without replacement from a box containing five red balls and ten blue balls. If $\overline{X}$ denotes the proportion of red balls in the sample, what are the mean and the variance of $\overline{X}$?

4. If a random variable X has the hypergeometric distribution with parameters $A = 8$, $B = 20$, and n, for what value of n will $\text{Var}(X)$ be a maximum?

5. Suppose that n students are selected at random without replacement from a class containing T students, of whom A are boys and $T - A$ are girls. Let X denote the number of boys that are obtained. For what sample size n will $\text{Var}(X)$ be a maximum?

6. Suppose that X_1 and X_2 are independent random variables, that X_1 has the binomial distribution with parameters n_1 and p, and that X_2 has the binomial distribution with parameters n_2 and p, where p is the same for both X_1 and X_2. For each fixed value of k ($k = 1, 2, \ldots, n_1 + n_2$), prove that the conditional distribution of X_1 given that $X_1 + X_2 = k$ is hypergeometric with parameters n_1, n_2, and k.

7. Suppose that in a large lot containing T manufactured items, 30 percent of the items are defective and 70 percent are nondefective. Also, suppose that ten items are selected at random without replacement from the lot. Determine **(a)** an exact expression for the probability that not more than one defective item will be obtained and **(b)** an approximate expression for this probability based on the binomial distribution.

8. Consider a group of T persons, and let $a_1, \ldots, a_T$ denote the heights of these T persons. Suppose that n persons are selected from this group at random without replacement, and let X denote the sum of the heights of these n persons. Determine the mean and variance of X.

9. Find the value of $\binom{3/2}{4}$.

10. Show that for all positive integers n and k,

$$\binom{-n}{k} = (-1)^k \binom{n+k-1}{k}.$$

11. Prove Theorem 5.3.3. *Hint:* Prove that

$$\lim_{n\to\infty} c_n \log(1 + a_n) - a_n c_n = 0$$

by applying Taylor's theorem with remainder (see Exercise 13 in Sec. 4.2) to the function $f(x) = \log(1 + x)$ around $x = 0$.

5.4 The Poisson Distributions

Many experiments consist of observing the occurrence times of random arrivals. Examples include arrivals of customers for service, arrivals of calls at a switchboard, occurrences of floods and other natural and man-made disasters, and so forth. The family of Poisson distributions is used to model the number of such arrivals that occur in a fixed time period. Poisson distributions are also useful approximations to binomial distributions with very small success probabilities.

Definition and Properties of the Poisson Distributions

Example 5.4.1

Customer Arrivals. A store owner believes that customers arrive at his store at a rate of 4.5 customers per hour on average. He wants to find the distribution of the actual number X of customers who will arrive during a particular one-hour period later in the day. He models customer arrivals in different time periods as independent of each other. As a first approximation, he divides the one-hour period into 3600 seconds and thinks of the arrival rate as being $4.5/3600 = 0.00125$ per second. He then says that during each second either 0 or 1 customers will arrive, and the probability of an arrival during any single second is 0.00125. He then tries to use the binomial distribution with

parameters $n = 3600$ and $p = 0.00125$ for the distribution of the number of customers who arrive during the one-hour period later in the day.

He starts calculating f, the p.f. of this binomial distribution, and quickly discovers how cumbersome the calculations are. However, he realizes that the successive values of $f(x)$ are closely related to each other because $f(x)$ changes in a systematic way as x increases. So he computes

$$\frac{f(x+1)}{f(x)} = \frac{\binom{n}{x+1}p^{x+1}(1-p)^{n-x-1}}{\binom{n}{x}p^x(1-p)^{n-x}} = \frac{(n-x)p}{(x+1)(1-p)} \approx \frac{np}{x+1},$$

where the reasoning for the approximation at the end is as follows: For the first 30 or so values of x, $n - x$ is essentially the same as n and dividing by $1 - p$ has almost no effect because p is so small. For example, for $x = 30$, the actual value is 0.1441, while the approximation is 0.1452. This approximation suggests defining $\lambda = np$ and approximating $f(x+1) \approx f(x)\lambda/(x+1)$ for all the values of x that matter. That is,

$$\begin{aligned} f(1) &= f(0)\lambda, \\ f(2) &= f(1)\frac{\lambda}{2} = f(0)\frac{\lambda^2}{2}, \\ f(3) &= f(2)\frac{\lambda}{3} = f(0)\frac{\lambda^3}{6}, \\ &\vdots \end{aligned}$$

Continuing the pattern for all x yields $f(x) = f(0)\lambda^x/x!$ for all x. To obtain a p.f. for X, he would need to make sure that $\sum_{x=0}^{\infty} f(x) = 1$. This is easily achieved by setting

$$f(0) = \frac{1}{\sum_{x=0}^{\infty} \lambda^x/x!} = e^{-\lambda},$$

where the last equality follows from the following well-known calculus result:

$$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}, \tag{5.4.1}$$

for all $\lambda > 0$. Hence, $f(x) = e^{-\lambda}\lambda^x/x!$ for $x = 0, 1, \ldots$ and $f(x) = 0$ otherwise is a p.f. ◀

The approximation formula for the p.f. of a binomial distribution at the end of Example 5.4.1 is actually a useful p.f. that can model many phenomena of types similar to the arrivals of customers.

Definition 5.4.1 Poisson Distribution. Let $\lambda > 0$. A random variable X has the *Poisson distribution with mean* λ if the p.f. of X is as follows:

$$f(x|\lambda) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!} & \text{for } x = 0, 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases} \tag{5.4.2}$$

At the end of Example 5.4.1, we proved that the function in Eq. (5.4.2) is indeed a p.f. In order to justify the phrase "with mean λ" in the definition of the distribution, we need to prove that the mean is indeed λ.

Theorem 5.4.1 Mean. The mean of the distribution with p.f. equal to (5.4.2) is λ.

Proof If X has the distribution with p.f. $f(x|\lambda)$, then $E(X)$ is given by the following infinite series:

$$E(X) = \sum_{x=0}^{\infty} x f(x|\lambda).$$

Since the term corresponding to $x = 0$ in this series is 0, we can omit this term and can begin the summation with the term for $x = 1$. Therefore,

$$E(X) = \sum_{x=1}^{\infty} x f(x|\lambda) = \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!}.$$

If we now let $y = x - 1$ in this summation, we obtain

$$E(X) = \lambda \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!}.$$

The sum of the series in this equation is the sum of $f(y|\lambda)$, which equals 1. Hence, $E(X) = \lambda$. ■

Example 5.4.2

Customer Arrivals. In Example 5.4.1, the store owner was approximating the binomial distribution with parameters 3600 and 0.00125 with a distribution that we now know as the Poisson distribution with mean $\lambda = 3600 \times 0.00125 = 4.5$. For $x = 0, \ldots, 9$, Table 5.1 has the binomial and corresponding Poisson probabilities.

The division of the one-hour period into 3600 seconds was somewhat arbitrary. The owner could have divided the hour into 7200 half-seconds or 14400 quarter-seconds, etc. Regardless of how finely the time is divided, the product of the number of time intervals and the rate in customers per time interval will always be 4.5 because they are all based on a rate of 4.5 customers per hour. Perhaps the store owner would do better simply modeling the number X of arrivals as a Poisson random variable with mean 4.5, rather than choosing an arbitrarily sized time interval to accommodate a tedious binomial calculation. The disadvantage to the Poisson model for X is that there is positive probability that a Poisson random variable will be arbitrarily large, whereas a binomial random variable with parameters n and p can never exceed n. However, the probability is essentially 0 that a Poisson random variable with mean 4.5 will exceed 19. ◄

Table 5.1 Binomial and Poisson probabilities in Example 5.4.2

	x				
	0	1	2	3	4
Binomial	0.01108	0.04991	0.11241	0.16874	0.18991
Poisson	0.01111	0.04999	0.11248	0.16872	0.18981
	x				
	5	6	7	8	9
Binomial	0.17094	0.12819	0.08237	0.04630	0.02313
Poisson	0.17083	0.12812	0.08237	0.04633	0.02317

Theorem 5.4.2 Variance. The variance of the Poisson distribution with mean λ is also λ.

Proof The variance can be found by a technique similar to the one used in the proof of Theorem 5.4.1 to find the mean. We begin by considering the following expectation:

$$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1) f(x|\lambda) = \sum_{x=2}^{\infty} x(x-1) f(x|\lambda)$$
$$= \sum_{x=2}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^x}{x!} = \lambda^2 \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^{x-2}}{(x-2)!}.$$

If we let $y = x - 2$, we obtain

$$E[X(X-1)] = \lambda^2 \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} = \lambda^2. \tag{5.4.3}$$

Since $E[X(X-1)] = E(X^2) - E(X) = E(X^2) - \lambda$, it follows from Eq. (5.4.3) that $E(X^2) = \lambda^2 + \lambda$. Therefore,

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \lambda. \tag{5.4.4}$$

Hence, the variance is also equal to λ. ■

Theorem 5.4.3 Moment Generating Function. The m.g.f. of the Poisson distribution with mean λ is

$$\psi(t) = e^{\lambda(e^t - 1)}, \tag{5.4.5}$$

for all real t.

Proof For every value of t $(-\infty < t < \infty)$,

$$\psi(t) = E(e^{tX}) = \sum_{x=0}^{\infty} \frac{e^{tx} e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!}.$$

It follows from Eq. (5.4.1) that, for $-\infty < t < \infty$,

$$\psi(t) = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}.$$ ■

The mean and the variance, as well as all other moments, can be determined from the m.g.f. given in Eq. (5.4.5). We shall not derive the values of any other moments here, but we shall use the m.g.f. to derive the following property of Poisson distributions.

Theorem 5.4.4 If the random variables $X_1, \dots, X_k$ are independent and if X_i has the Poisson distribution with mean λ_i $(i = 1, \dots, k)$, then the sum $X_1 + \cdots + X_k$ has the Poisson distribution with mean $\lambda_1 + \cdots + \lambda_k$.

Proof Let $\psi_i(t)$ denote the m.g.f. of X_i for $i = 1, \dots, k$, and let $\psi(t)$ denote the m.g.f. of the sum $X_1 + \cdots + X_k$. Since $X_1, \dots, X_k$ are independent, it follows that, for $-\infty < t < \infty$,

$$\psi(t) = \prod_{i=1}^{k} \psi_i(t) = \prod_{i=1}^{k} e^{\lambda_i(e^t - 1)} = e^{(\lambda_1 + \cdots + \lambda_k)(e^t - 1)}.$$

It can be seen from Eq. (5.4.5) that this m.g.f. $\psi(t)$ is the m.g.f. of the Poisson distribution with mean $\lambda_1 + \cdots + \lambda_k$. Hence, the distribution of $X_1 + \cdots + X_k$ must be as stated in the theorem. ■

A table of probabilities for Poisson distributions with various values of the mean λ is given at the end of this book.

Example 5.4.3 Customer Arrivals. Suppose that the store owner in Examples 5.4.1 and 5.4.2 is interested not only in the number of customers that arrive in the one-hour period, but also in how many customers arrive in the next hour after that period. Let Y be the number of customers that arrive in the second hour. By the reasoning at the end of Example 5.4.2, the owner might model Y as a Poisson random variable with mean 4.5. He would also say that X and Y are independent because he has been assuming that arrivals in disjoint time intervals are independent. According to Theorem 5.4.4, $X + Y$ would have the Poisson distribution with mean $4.5 + 4.5 = 9$. What is the probability that at least 12 customers will arrive in the entire two-hour period? We can use the table of Poisson probabilities in the back of this book by looking in the $\lambda = 9$ column. Either add up the numbers corresponding to $k = 0, \ldots, 11$ and subtract the total from 1, or add up those from $k = 12$ to the end. Either way, the result is $\Pr(X \geq 12) = 0.1970$. ◄

The Poisson Approximation to Binomial Distributions

In Examples 5.4.1 and 5.4.2, we illustrated how close the Poisson distribution with mean 4.5 is to the binomial distribution with parameters 3600 and 0.00125. We shall now demonstrate a general version of that result, namely, that when the value of n is large and the value of p is close to 0, the binomial distribution with parameters n and p can be approximated by the Poisson distribution with mean np.

Theorem 5.4.5 Closeness of Binomial and Poisson Distributions. For each integer n and each $0 < p < 1$, let $f(x|n, p)$ denote the p.f. of the binomial distribution with parameters n and p. Let $f(x|\lambda)$ denote the p.f. of the Poisson distribution with mean λ. Let $\{p_n\}_{n=1}^{\infty}$ be a sequence of numbers between 0 and 1 such that $\lim_{n\to\infty} np_n = \lambda$. Then

$$\lim_{n\to\infty} f(x|n, p_n) = f(x|\lambda),$$

for all $x = 0, 1, \ldots$.

Proof We begin by writing

$$f(x|n, p_n) = \frac{n(n-1)\cdots(n-x+1)}{x!} p_n^x (1-p_n)^{n-x}.$$

Next, let $\lambda_n = np_n$ so that $\lim_{n\to\infty} \lambda_n = \lambda$. Then $f(x|n, p_n)$ can be rewritten in the following form:

$$f(x|n, p_n) = \frac{\lambda_n^x}{x!}\frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-x+1}{n}\left(1 - \frac{\lambda_n}{n}\right)^n \left(1 - \frac{\lambda_n}{n}\right)^{-x}. \quad (5.4.6)$$

For each $x \geq 0$,

$$\lim_{n\to\infty} \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-x+1}{n}\left(1 - \frac{\lambda_n}{n}\right)^{-x} = 1.$$

Furthermore, it follows from Theorem 5.3.3 that

$$\lim_{n\to\infty}\left(1-\frac{\lambda_n}{n}\right)^n = e^{-\lambda}. \tag{5.4.7}$$

It now follows from Eq. (5.4.6) that for every $x \ge 0$,

$$\lim_{n\to\infty} f(x|n,\, p_n) = \frac{e^{-\lambda}\lambda^x}{x!} = f(x|\lambda). \qquad \blacksquare$$

Example 5.4.4 Approximating a Probability. Suppose that in a large population the proportion of people who have a certain disease is 0.01. We shall determine the probability that in a random group of 200 people at least four people will have the disease.

In this example, we can assume that the exact distribution of the number of people having the disease among the 200 people in the random group is the binomial distribution with parameters $n = 200$ and $p = 0.01$. Therefore, this distribution can be approximated by the Poisson distribution for which the mean is $\lambda = np = 2$. If X denotes a random variable having this Poisson distribution, then it can be found from the table of the Poisson distribution at the end of this book that $\Pr(X \ge 4) = 0.1428$. Hence, the probability that at least four people will have the disease is approximately 0.1428. The actual value is 0.1420. ◀

Theorem 5.4.5 says that if n is large and p is small so that np is close to λ, then the binomial distribution with parameters n and p is close to the Poisson distribution with mean λ. Recall Theorem 5.3.4, which says that if A and B are large compared to n and if $A/(A+B)$ is close to p, then the hypergeometric distribution with parameters A, B, and n is close to the binomial distribution with parameters n and p. These two results can be combined into the following theorem, whose proof is left to Exercise 17.

Theorem 5.4.6 Closeness of Hypergeometric and Poisson Distributions. Let $\lambda > 0$. Let Y have the Poisson distribution with mean λ. For each positive integer T, let A_T, B_T, and n_T be integers such that $\lim_{T\to\infty} A_T = \infty$, $\lim_{T\to\infty} B_T = \infty$, $\lim_{T\to\infty} n_T = \infty$, and $\lim_{T\to\infty} n_T A_T/(A_T + B_T) = \lambda$. Let X_T have the hypergeometric distribution with parameters A_T, B_T, and n_T. For each fixed $x = 0, 1, \ldots$,

$$\lim_{T\to\infty} \frac{\Pr(Y = x)}{\Pr(X_T = x)} = 1. \qquad \blacksquare$$

Poisson Processes

Example 5.4.5 Customer Arrivals. In Example 5.4.3, the store owner believes that the number of customers that arrive in each one-hour period has the Poisson distribution with mean 4.5. What if the owner is interested in a half-hour period or a 4-hour and 15-minute period? Is it safe to assume that the number of customers that arrive in a half-hour period has the Poisson distribution with mean 2.25? ◀

In order to be sure that all of the distributions for the various numbers of arrivals in Example 5.4.5 are consistent with each other, the store owner needs to think about the overall process of customer arrivals, not just a few isolated time periods. The following definition gives a model for the overall process of arrivals that will allow the store owner to construct distributions for all the counts of customer arrivals that interest him as well as other useful things.

Definition 5.4.2 Poisson Process. A *Poisson process* with rate λ per unit time is a process that satisfies the following two properties:

i. The number of arrivals in every fixed interval of time of length t has the Poisson distribution with mean λt.
ii. The numbers of arrivals in every collection of disjoint time intervals are independent.

The answer to the question at the end of Example 5.4.5 will be "yes" if the store owner makes the assumption that customers arrive according to a Poisson process with rate 4.5 per hour. Here is another example.

Example 5.4.6 Radioactive Particles. Suppose that radioactive particles strike a certain target in accordance with a Poisson process at an average rate of three particles per minute. We shall determine the probability that 10 or more particles will strike the target in a particular two-minute period.

In a Poisson process, the number of particles striking the target in any particular one-minute period has the Poisson distribution with mean λ. Since the mean number of strikes in any one-minute period is 3, it follows that $\lambda = 3$ in this example. Therefore, the number of strikes X in any two-minute period will have the Poisson distribution with mean 6. It can be found from the table of the Poisson distribution at the end of this book that $\Pr(X \geq 10) = 0.0838$. ◄

Note: Generality of Poisson Processes. Although we have introduced Poisson processes in terms of counts of arrivals during time intervals, Poisson processes are actually more general. For example, a Poisson process can be used to model occurrences in space as well as time. A Poisson process could be used to model telephone calls arriving at a switchboard, atomic particles emitted from a radioactive source, diseased trees in a forest, or defects on the surface of a manufactured product. The reason for the popularity of the Poisson process model is twofold. First, the model is computationally convenient. Second, there is a mathematical justification for the model if one makes three plausible assumptions about how the phenomena occur. We shall present the three assumptions in some detail after another example.

Example 5.4.7 Cryptosporidium in Drinking Water. Cryptosporidium is a genus of protozoa that occurs as small oocysts and can cause painful sickness and even death when ingested. Occasionally, oocysts are detected in public drinking water supplies. A concentration as low as one oocyst per five liters can be enough to trigger a boil-water advisory. In April 1993, many thousands of people became ill during a cryptosporidiosis outbreak in Milwaukee, Wisconsin. Different water systems have different systems for monitoring protozoa occurrence in drinking water. One problem with monitoring systems is that detection technology is not always very sensitive. One popular technique is to push a large amount of water through a very fine filter and then treat the material captured on the filter in a way that identifies Cryptosporidium oocysts. The number of oocysts is then counted and recorded. Even if there is an oocyst on the filter, the probability can be as low as 0.1 that it will get counted.

Suppose that, in a particular water supply, oocysts occur according to a Poisson process with rate λ oocysts per liter. Suppose that the filtering system is capable of capturing all oocysts in a sample, but that the counting system has probability p of actually observing each oocyst that is on the filter. Assume that the counting system observes or misses each oocyst on the filter independently. What is the distribution of the number of counted oocysts from t liters of filtered water?

Let Y be the number of oocysts in the t liters (all of which make it onto the filter). Then Y has the Poisson distribution with mean λt. Let $X_i = 1$ if the ith oocyst on the filter gets counted, and $X_i = 0$ if not. Let X be the counted number of oocysts so that $X = X_1 + \cdots + X_y$ if $Y = y$. Conditional on $Y = y$, we have assumed that the X_i are independent Bernoulli random variables with parameter p, so X has the binomial distribution with parameters y and p conditional on $Y = y$. We want the marginal distribution of X. This can be found using the law of total probability for random variables (3.6.11). For $x = 0, 1, \ldots,$

$$\begin{aligned}
f_1(x) &= \sum_{y=0}^{\infty} g_1(x|y) f_2(y) \\
&= \sum_{y=x}^{\infty} \binom{y}{x} p^x (1-p)^{y-x} e^{-\lambda t} \frac{(\lambda t)^y}{y!} \\
&= e^{-\lambda t} \frac{(p\lambda t)^x}{x!} \sum_{y=x}^{\infty} \frac{[\lambda t(1-p)]^{y-x}}{(y-x)!} \\
&= e^{-\lambda t} \frac{(p\lambda t)^x}{x!} \sum_{u=0}^{\infty} \frac{[\lambda t(1-p)]^{u}}{u!} \\
&= e^{-\lambda t} \frac{(p\lambda t)^x}{x!} e^{\lambda t(1-p)} = e^{-p\lambda t} \frac{(p\lambda t)^x}{x!}.
\end{aligned}$$

This is easily recognized as the p.f. of the Poisson distribution with mean $p\lambda t$. The effect of losing a fraction $1 - p$ of the oocyst count is merely to lower the rate of the Poisson process from λ per liter to $p\lambda$ per liter.

Suppose that $\lambda = 0.2$ and $p = 0.1$. How much water must we filter in order for there to be probability at least 0.9 that we will count at least one oocyst? The probability of counting at least one oocyst is 1 minus the probability of counting none, which is $e^{-p\lambda t} = e^{-0.02t}$. So, we need t large enough so that $1 - e^{-0.02t} \geq 0.9$, that is, $t \geq 115$. A typical procedure is to test 100 liters, which would have probability $1 - e^{-.02\times 100} = 0.86$ of detecting at least one oocyst. ◄

Assumptions Underlying the Poisson Process Model

In what follows, we shall refer to time intervals, but the assumptions can be used equally well for subregions of two- or three-dimensional regions or sublengths of a linear distance. Indeed, a Poisson process can be used to model occurrences in any region that can be subdivided into arbitrarily small pieces. There are three assumptions that lead to the Poisson process model.

The first assumption is that the numbers of occurrences in any collection of *disjoint* intervals of time must be mutually independent. For example, even though an unusually large number of telephone calls are received at a switchboard during a particular interval, the probability that at least one call will be received during a forthcoming interval remains unchanged. Similarly, even though no call has been received at the switchboard for an unusually long interval, the probability that a call will be received during the next short interval remains unchanged.

The second assumption is that the probability of an occurrence during each very short interval of time must be approximately proportional to the length of that interval. To express this condition more formally, we shall use the standard

mathematical notation in which $o(t)$ denotes any function of t having the property that

$$\lim_{t \to 0} \frac{o(t)}{t} = 0. \tag{5.4.8}$$

According to (5.4.8), $o(t)$ must be a function that approaches 0 as $t \to 0$, and, furthermore, this function must approach 0 at a rate faster than t itself. An example of such a function is $o(t) = t^{\alpha}$, where $\alpha > 1$. It can be verified that this function satisfies Eq. (5.4.8). The second assumption can now be expressed as follows: There exists a constant $\lambda > 0$ such that for every time interval of length t, the probability of at least one occurrence during that interval has the form $\lambda t + o(t)$. Thus, for every very small value of t, the probability of at least one occurrence during an interval of length t is equal to λt plus a quantity having a smaller order of magnitude.

One of the consequences of the second assumption is that the process being observed must be *stationary* over the entire period of observation; that is, the probability of an occurrence must be the same over the entire period. There can be neither busy intervals, during which we know in advance that occurrences are likely to be more frequent, nor quiet intervals, during which we know in advance that occurrences are likely to be less frequent. This condition is reflected in the fact that the same constant λ expresses the probability of an occurrence in every interval over the entire period of observation. The second assumption can be relaxed at the cost of more complicated mathematics, but we shall not do so here.

The third assumption is that, for each very short interval of time, the probability that there will be two or more occurrences in that interval must have a smaller order of magnitude than the probability that there will be just one occurrence. In symbols, the probability of two or more occurrences in a time interval of length t must be $o(t)$. Thus, the probability of two or more occurrences in a small interval must be negligible in comparison with the probability of one occurrence in that interval. Of course, it follows from the second assumption that the probability of one occurrence in that same interval will itself be negligible in comparison with the probability of no occurrences.

Under the preceding three assumptions, it can be shown that the process will satisfy the definition of a Poisson process with rate λ. See Exercise 16 in this section for one method of proof.

Summary

Poisson distributions are used to model data that arrive as counts. A Poisson process with rate λ is a model for random occurrences that have a constant expected rate λ per unit time (or per unit area). We must assume that occurrences in disjoint time intervals (or disjoint areas) are independent and that two or more occurrences cannot happen at the same time (or place). The number of occurrences in an interval of length (or area of size) t has the Poisson distribution with mean $t\lambda$. If n is large and p is small, then the binomial distribution with parameters n and p is approximately the same as the Poisson distribution with mean np.

Exercises

1. In Example 5.4.7, with $\lambda = 0.2$ and $p = 0.1$, compute the probability that we would detect at least two oocysts after filtering 100 liters of water.

2. Suppose that on a given weekend the number of accidents at a certain intersection has the Poisson distribution with mean 0.7. What is the probability that there will be at least three accidents at the intersection during the weekend?

3. Suppose that the number of defects on a bolt of cloth produced by a certain process has the Poisson distribution with mean 0.4. If a random sample of five bolts of cloth is inspected, what is the probability that the total number of defects on the five bolts will be at least 6?

4. Suppose that in a certain book there are on the average λ misprints per page and that misprints occurred according to a Poisson process. What is the probability that a particular page will contain no misprints?

5. Suppose that a book with n pages contains on the average λ misprints per page. What is the probability that there will be at least m pages which contain more than k misprints?

6. Suppose that a certain type of magnetic tape contains on the average three defects per 1000 feet. What is the probability that a roll of tape 1200 feet long contains no defects?

7. Suppose that on the average a certain store serves 15 customers per hour. What is the probability that the store will serve more than 20 customers in a particular two-hour period?

8. Suppose that X_1 and X_2 are independent random variables and that X_i has the Poisson distribution with mean λ_i $(i = 1, 2)$. For each fixed value of k $(k = 1, 2, \ldots)$, determine the conditional distribution of X_1 given that $X_1 + X_2 = k$.

9. Suppose that the total number of items produced by a certain machine has the Poisson distribution with mean λ, all items are produced independently of one another, and the probability that any given item produced by the machine will be defective is p. Determine the marginal distribution of the number of defective items produced by the machine.

10. For the problem described in Exercise 9, let X denote the number of defective items produced by the machine, and let Y denote the number of nondefective items produced by the machine. Show that X and Y are independent random variables.

11. The mode of a discrete distribution was defined in Exercise 12 of Sec. 5.2. Determine the mode or modes of the Poisson distribution with mean λ.

12. Suppose that the proportion of colorblind people in a certain population is 0.005. What is the probability that there will not be more than one colorblind person in a randomly chosen group of 600 people?

13. The probability of triplets in human births is approximately 0.001. What is the probability that there will be exactly one set of triplets among 700 births in a large hospital?

14. An airline sells 200 tickets for a certain flight on an airplane that has only 198 seats because, on the average, 1 percent of purchasers of airline tickets do not appear for the departure of their flight. Determine the probability that everyone who appears for the departure of this flight will have a seat.

15. Suppose that internet users access a particular Web site according to a Poisson process with rate λ per hour, but λ is unknown. The Web site maintainer believes that λ has a continuous distribution with p.d.f.

$$f(\lambda) = \begin{cases} 2e^{-2\lambda} & \text{for } \lambda > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let X be the number of users who access the Web site during a one-hour period. If $X = 1$ is observed, find the conditional p.d.f. of λ given $X = 1$.

16. In this exercise, we shall prove that the three assumptions underlying the Poisson process model do indeed imply that occurrences happen according to a Poisson process. What we need to show is that, for each t, the number of occurrences during a time interval of length t has the Poisson distribution with mean λt. Let X stand for the number of occurrences during a particular time interval of length t. Feel free to use the following extension of Eq. (5.4.7): For all real a,

$$\lim_{u \to 0} (1 + au + o(u))^{1/u} = e^a, \tag{5.4.9}$$

a. For each positive integer n, divide the time interval into n disjoint subintervals of length t/n each. For $i = 1, \ldots, n$, let $Y_i = 1$ if exactly one arrival occurs in the ith subinterval, and let A_i be the event that two or more occurrences occur during the ith subinterval. Let $W_n = \sum_{i=1}^n Y_i$. For each nonnegative integer k, show that we can write $\Pr(X = k) = \Pr(W_n = k) + \Pr(B)$, where B is a subset of $\cup_{i=1}^n A_i$.

b. Show that $\lim_{n\to\infty} \Pr(\cup_{i=1}^n A_i) = 0$. *Hint:* Show that $\Pr(\cap_{i=1}^n A_i^c) = (1 + o(u))^{1/u}$ where $u = 1/n$.

c. Show that $\lim_{n\to\infty} \Pr(W_n = k) = e^{-\lambda}(\lambda t)^k/k!$. *Hint:* $\lim_{n\to\infty} n!/[n^k(n-k)!] = 1$.

d. Show that X has the Poisson distribution with mean λt.

17. Prove Theorem 5.4.6. One approach is to adapt the proof of Theorem 5.3.4 by replacing n by n_T in that proof. The steps of the proof that are significanlty different are the following. (i) You will need to show that $B_T - n_T$ goes to ∞. (ii) The three limits that depend on Theorem 5.3.3 need to be rewritten as ratios converging to 1. For example, the second one is rewritten as

$$\lim_{T\to\infty}\left(\frac{B_T}{B_T - n_T + x}\right)^{B_T - n_T + x + 1/2} e^{-n_T + x} = 1.$$

You'll need a couple more such limits as well. (iii) Instead of (5.3.12), prove that

$$\lim_{T\to\infty}\frac{n_T^x A_T^x B_T^{n_T - x}}{(A_T + B_T)^{n_T}} = \lambda^x e^{-\lambda}.$$

18. Let A_T, B_T, and n_T be sequences, all three of which go to ∞ as $T \to \infty$. Prove that $\lim_{T\to\infty} n_T A_T/(A_T + B_T) = \lambda$ if and only if $\lim_{T\to\infty} n_T A_T/B_T = \lambda$.

5.5 The Negative Binomial Distributions

Earlier we learned that, in n Bernoulli trials with probability of success p, the number of successes has the binomial distribution with parameters n and p. Instead of counting successes in a fixed number of trials, it is often necessary to observe the trials until we see a fixed number of successes. For example, while monitoring a piece of equipment to see when it needs maintenance, we might let it run until it produces a fixed number of errors and then repair it. The number of failures until a fixed number of successes has a distribution in the family of negative binomial distributions.

Definition and Interpretation

Example 5.5.1 Defective Parts. Suppose that a machine produces parts that can be either good or defective. Let $X_i = 1$ if the ith part is defective and $X_i = 0$ otherwise. Assume that the parts are good or defective independently of each other with $\Pr(X_i = 1) = p$ for all i. An inspector observes the parts produced by this machine until she sees four defectives. Let X be the number of good parts observed by the time that the fourth defective is observed. What is the distribution of X? ◀

The problem described in Example 5.5.1 is typical of a general situation in which a sequence of Bernoulli trials can be observed. Suppose that an infinite sequence of Bernoulli trials is available. Call the two possible outcomes success and failure, with p being the probability of success. In this section, we shall study the distribution of the total number of failures that will occur before exactly r successes have been obtained, where r is a fixed positive integer.

Theorem 5.5.1 Sampling until a Fixed Number of Successes. Suppose that an infinite sequence of Bernoulli trials with probability of success p are available. The number X of failures that occur before the rth success has the following p.d.f.:

$$f(x|r, p) = \begin{cases} \binom{r + x - 1}{x} p^r (1-p)^x & \text{for } x = 0, 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases} \quad (5.5.1)$$

Proof For $n = r, r + 1, \ldots$, we shall let A_n denote the event that the total number of trials required to obtain exactly r successes is n. As explained in Example 2.2.8, the event A_n will occur if and only if exactly $r - 1$ successes occur among the first $n - 1$

trials and the rth success is obtained on the nth trial. Since all trials are independent, it follows that

$$\Pr(A_n) = \binom{n-1}{r-1} p^{r-1}(1-p)^{(n-1)-(r-1)} \cdot p = \binom{n-1}{r-1} p^r (1-p)^{n-r}. \quad (5.5.2)$$

For each value of x ($x = 0, 1, 2, \ldots$), the event that exactly x failures are obtained before the rth success is obtained is the same as the event that the total number of trials required to obtain r successes is $r + x$. In other words, if X denotes the number of failures that will occur before the rth success is obtained, then $\Pr(X = x) = \Pr(A_{r+x})$. Eq. (5.5.1) now follows from Eq. (5.5.2). ■

Definition 5.5.1 Negative Binomial Distribution. A random variable X has the *negative binomial distribution with parameters r and p* ($r = 1, 2, \ldots$ and $0 < p < 1$) if X has a discrete distribution for which the p.f. $f(x|r, p)$ is as specified by Eq. (5.5.1).

Example 5.5.2 Defective Parts. Example 5.5.1 is worded so that defective parts are successes and good parts are failures. The distribution of the number X of good parts observed by the time of the fourth defective is the negative binomial distribution with parameters 4 and p. ◄

The Geometric Distributions

The most common special case of a negative binomial random variable is one for which $r = 1$. This would be the number of failures until the first success.

Definition 5.5.2 Geometric Distribution. A random variable X has the *geometric distribution with parameter p* ($0 < p < 1$) if X has a discrete distribution for which the p.f. $f(x|1, p)$ is as follows:

$$f(x|1, p) = \begin{cases} p(1-p)^x & \text{for } x = 0, 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases} \quad (5.5.3)$$

Example 5.5.3 Triples in the Lottery. A common daily lottery game involves the drawing of three digits from 0 to 9 independently with replacement and independently from day to day. Lottery watchers often get excited when all three digits are the same, an event called *triples*. If p is the probability of obtaining triples, and if X is the number of days without triples before the first triple is observed, then X has the geometric distribution with parameter p. In this case, it is easy to see that $p = 0.01$, since there are 10 different triples among the 1000 equally likely daily numbers. ◄

The relationship between geometric and negative binomial distributions goes beyond the fact that the geometric distributions are special cases of negative binomial distributions.

Theorem 5.5.2 If $X_1, \ldots, X_r$ are i.i.d. random variables and if each X_i has the geometric distribution with parameter p, then the sum $X_1 + \cdots + X_r$ has the negative binomial distribution with parameters r and p.

Proof Consider an infinite sequence of Bernoulli trials with success probability p. Let X_1 denote the number of failures that occur before the first success is obtained; then X_1 will have the geometric distribution with parameter p.

Now continue observing the Bernoulli trials after the first success. For $j = 2, 3, \ldots$, let X_j denote the number of failures that occur after $j - 1$ successes have

been obtained but before the jth success is obtained. Since all the trials are independent and the probability of obtaining a success on each trial is p, it follows that each random variable X_j will have the geometric distribution with parameter p and that the random variables $X_1, X_2, \ldots$ will be independent. Furthermore, for $r = 1, 2, \ldots,$ the sum $X_1 + \cdots + X_r$ will be equal to the total number of failures that occur before exactly r successes have been obtained. Therefore, this sum will have the negative binomial distribution with parameters r and p. ■

Properties of Negative Binomial and Geometric Distributions

Theorem 5.5.3 Moment Generating Function. If X has the negative binomial distribution with parameters r and p, then the m.g.f. of X is as follows:

$$\psi(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r \quad \text{for } t < \log\left(\frac{1}{1-p}\right). \tag{5.5.4}$$

The m.g.f. of the geometric distribution with parameter p is the special case of Eq. (5.5.4) with $r = 1$.

Proof Let $X_1, \ldots, X_r$ be a random sample of r geometric random variables each with parameter p. We shall find the m.g.f. of X_1 and then apply Theorems 4.4.4 and 5.5.2 to find the m.g.f. of the negative binomial distribution with parameters r and p.

The m.g.f. $\psi_1(t)$ of X_1 is

$$\psi_1(t) = E(e^{tX_1}) = p\sum_{x=0}^{\infty}[(1-p)e^t]^x. \tag{5.5.5}$$

The infinite series in Eq. (5.5.5) will have a finite sum for every value of t such that $0 < (1-p)e^t < 1$, that is, for $t < \log(1/[1-p])$. It is known from elementary calculus that for every number α $(0 < \alpha < 1)$,

$$\sum_{x=0}^{\infty}\alpha^x = \frac{1}{1-\alpha}.$$

Therefore, for $t < \log(1/[1-p])$, the m.g.f. of the geometric distribution with parameter p is

$$\psi_1(t) = \frac{p}{1-(1-p)e^t}. \tag{5.5.6}$$

Each of $X_1, \ldots, X_r$ has the same m.g.f., namely, ψ_1. According to Theorem 4.4.4, the m.g.f. of $X = X_1 + \cdots + X_r$ is $\psi(t) = [\psi_1(t)]^r$. Theorem 5.5.2 says that X has the negative binomial distribution with parameters r and p, and hence the m.g.f. of X is $[\psi_1(t)]^r$, which is the same as Eq. (5.5.4). ■

Theorem 5.5.4 Mean and Variance. If X has the negative binomial distribution with parameters r and p, the mean and the variance of X must be

$$E(X) = \frac{r(1-p)}{p} \quad \text{and} \quad \text{Var}(X) = \frac{r(1-p)}{p^2}. \tag{5.5.7}$$

The mean and variance of the geometric distribution with parameter p are the special case of Eq. (5.5.7) with $r = 1$.

Proof Let X_1 have the geometric distribution with parameter p. We will find the mean and variance by differentiating the m.g.f. Eq. (5.5.5):

$$E(X_1) = \psi_1'(0) = \frac{1-p}{p}, \tag{5.5.8}$$

$$\mathrm{Var}(X_1) = \psi_1''(0) - [\psi_1'(0)]^2 = \frac{1-p}{p^2}. \tag{5.5.9}$$

If X has the negative binomial distribution with parameters r and p, represent it as the sum $X = X_1 + \cdots + X_r$ of r independent random variables, each having the same distribution as X_1. Eq. (5.5.7) now follows from Eqs. (5.5.8) and (5.5.9). ■

Example 5.5.4 Triples in the Lottery. In Example 5.5.3, the number X of daily draws without a triple until we see a triple has the geometric distribution with parameter $p = 0.01$. The total number of days until we see the first triple is then $X + 1$. So, the expected number of days until we observe triples is $E(X) + 1 = 100$.

Now suppose that a lottery player has been waiting 120 days for triples to occur. Such a player might conclude from the preceeding calculation that triples are "due." The most straightforward way to address such a claim would be to start by calculating the conditional distribution of X given that $X \geq 120$. ◀

The next result says that the lottery player at the end of Example 5.5.4 couldn't be farther from correct. Regardless of how long he has waited for triples, the time remaining until triples occur has the same geometric distribution (and the same mean) as it had when he started waiting. The proof is simple and is left as Exercise 8.

Theorem 5.5.5 Memoryless Property of Geometric Distributions. Let X have the geometric distribution with parameter p, and let $k \geq 0$. Then for every integer $t \geq 0$,

$$\Pr(X = k + t | X \geq k) = \Pr(X = t). \quad ■$$

The intuition behind Theorem 5.5.5 is the following: Think of X as the number of failures until the first success in a sequence of Bernoulli trials. Let Y be the number of failures starting with the $k + 1$st trial until the next success. Then Y has the same distribution as X and is independent of the first k trials. Hence, conditioning on anything that happened on the first k trials, such as no successes yet, doesn't affect the distribution of Y—it is still the same geometric distribution. A formal proof can be given in Exercise 8. In Exercise 13, you can prove that the geometric distributions are the only discrete distributions that have the memoryless property.

Example 5.5.5 Triples in the Lottery. In Example 5.5.4, after the first 120 non-triples, the process essentially starts over again and we still have to wait a geometric amount of time until the first triple.

At the beginning of the experiment, the expected number of failures (non-triples) that will occur before the first success (triples) is $(1 - p)/p$, as given by Eq. (5.5.8). If it is known that failures were obtained on the first 120 trials, then the conditional expected total number of failures before the first success (given the 120 failures on the first 120 trials) is simply $120 + (1 - p)/p$. ◀

Extension of Definition of Negative Binomial Distributon

By using the definition of binomial coefficients given in Eq. (5.3.14), the function $f(x|r, p)$ can be regarded as the p.f. of a discrete distribution for each number $r > 0$ (not necessarily an integer) and each number p in the interval $0 < p < 1$. In other words, it can be verified that for $r > 0$ and $0 < p < 1$,

$$\sum_{x=0}^{\infty} \binom{r + x - 1}{x} p^r (1 - p)^x = 1. \tag{5.5.10}$$

Summary

If we observe a sequence of independent Bernoulli trials with success probability p, the number of failures until the rth success has the negative binomial distribution with parameters r and p. The special case of $r = 1$ is the geometric distribution with parameter p. The sum of independent negative binomial random variables with the same second parameter p has a negative binomial distribution.

Exercises

1. Consider a daily lottery as described in Example 5.5.4.

a. Compute the probability that two particular days in a row will both have triples.

b. Suppose that we observe triples on a particular day. Compute the conditional probability that we observe triples again the next day.

2. Suppose that a sequence of independent tosses are made with a coin for which the probability of obtaining a head on each given toss is 1/30.

a. What is the expected number of tails that will be obtained before five heads have been obtained?

b. What is the variance of the number of tails that will be obtained before five heads have been obtained?

3. Consider the sequence of coin tosses described in Exercise 2.

a. What is the expected number of tosses that will be required in order to obtain five heads?

b. What is the variance of the number of tosses that will be required in order to obtain five heads?

4. Suppose that two players A and B are trying to throw a basketball through a hoop. The probability that player A will succeed on any given throw is p, and he throws until he has succeeded r times. The probability that player B will succeed on any given throw is mp, where m is a given integer ($m = 2, 3, \ldots$) such that $mp < 1$, and she throws until she has succeeded mr times.

a. For which player is the expected number of throws smaller?

b. For which player is the variance of the number of throws smaller?

5. Suppose that the random variables $X_1, \ldots, X_k$ are independent and that X_i has the negative binomial distribution with parameters r_i and p $(i = 1 \ldots k)$. Prove that the sum $X_1 + \cdots + X_k$ has the negative binomial distribution with parameters $r = r_1 + \cdots + r_k$ and p.

6. Suppose that X has the geometric distribution with parameter p. Determine the probability that the value of X will be one of the even integers $0, 2, 4, \ldots$.

7. Suppose that X has the geometric distribution with parameter p. Show that for every nonnegative integer k, $\Pr(X \geq k) = (1 - p)^k$.

8. Prove Theorem 5.5.5.

9. Suppose that an electronic system contains n components that function independently of each other, and suppose that these components are connected in series, as defined in Exercise 5 of Sec. 3.7. Suppose also that each component will function properly for a certain number of periods and then will fail. Finally, suppose that for $i = 1, \ldots, n$, the number of periods for which component i will function properly is a discrete random variable having

a geometric distribution with parameter p_i. Determine the distribution of the number of periods for which the system will function properly.

10. Let $f(x|r, p)$ denote the p.f. of the negative binomial distribution with parameters r and p, and let $f(x|\lambda)$ denote the p.f. of the Poisson distribution with mean λ, as defined by Eq. (5.4.2). Suppose $r \to \infty$ and $p \to 1$ in such a way that the value of $r(1-p)$ remains constant and is equal to λ throughout the process. Show that for each fixed nonnegative integer x,

$$f(x|r, p) \to f(x|\lambda).$$

11. Prove that the p.f. of the negative binomial distribution can be written in the following alternative form:

$$f(x|r, p) = \begin{cases} \binom{-r}{x} p^r(-[1-p])^x & \text{for } x = 0, 1, 2, \dots, \\ 0 & \text{otherwise.} \end{cases}$$

Hint: Use Exercise 10 in Sec. 5.3.

12. Suppose that a machine produces parts that are defective with probability P, but P is unknown. Suppose that P has a continuous distribution with p.d.f.

$$f(p) = \begin{cases} 10(1-p)^9 & \text{if } 0 < p < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Conditional on $P = p$, assume that all parts are independent of each other. Let X be the number of nondefective parts observed until the first defective part. If we observe $X = 12$, compute the conditional p.d.f. of P given $X = 12$.

13. Let F be the c.d.f. of a discrete distribution that has the memoryless property stated in Theorem 5.5.5. Define $\ell(x) = \log[1 - F(x-1)]$ for $x = 1, 2, \dots$.

a. Show that, for all integers $t, h > 0$,

$$1 - F(h-1) = \frac{1 - F(t+h-1)}{1 - F(t-1)}.$$

b. Prove that $\ell(t+h) = \ell(t) + \ell(h)$ for all integers $t, h > 0$.

c. Prove that $\ell(t) = t\ell(1)$ for every integer $t > 0$.

d. Prove that F must be the c.d.f. of a geometric distribution.

5.6 The Normal Distributions

The most widely used model for random variables with continuous distributions is the family of normal distributions. These distributions are the first ones we shall see whose p.d.f.'s cannot be integrated in closed form, and hence tables of the c.d.f. or computer programs are necessary in order to compute probabilities and quantiles for normal distributions.

Importance of the Normal Distributions

Example 5.6.1

Automobile Emissions. Automobile engines emit a number of undesirable pollutants when they burn gasoline. Lorenzen (1980) studied the amounts of various pollutants emitted by 46 automobile engines. One class of polutants consists of the oxides of nitrogen. Figure 5.1 shows a histogram of the 46 amounts of oxides of nitrogen (in grams per mile) that are reported by Lorenzen (1980). The bars in the histogram have areas that equal the proportions of the sample of 46 measurements that lie between the points on the horizontal axis where the sides of the bars stand. For example, the fourth bar (which runs from 1.0 to 1.2 on the horizontal axis) has area $0.870 \times 0.2 = 0.174$, which equals 8/46 because there are eight observations between 1.0 and 1.2. When we want to make statements about probabilities related to emissions, we will need a distribution with which to model emissions. The family of normal distributions introduced in this section will prove to be valuable in examples such as this. ◄

The family of normal distributions, which will be defined and discussed in this section, is by far the single most important collection of probability distributions

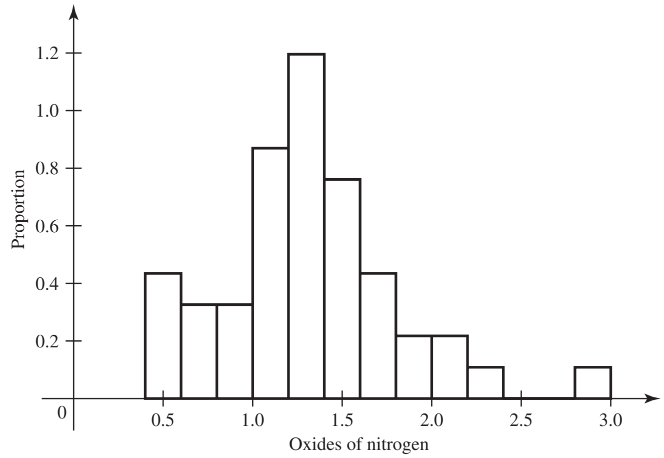


Figure 5.1 Histogram of emissions of oxides of nitrogen for Example 5.6.1 in grams per mile over a common driving regimen.

in statistics. There are three main reasons for this preeminent position of these distributions.

The first reason is directly related to the mathematical properties of the normal distributions. We shall demonstrate in this section and in several later sections of this book that if a random sample is taken from a normal distribution, then the distributions of various important functions of the observations in the sample can be derived explicitly and will themselves have simple forms. Therefore, it is a mathematical convenience to be able to assume that the distribution from which a random sample is drawn is a normal distribution.

The second reason is that many scientists have observed that the random variables studied in various physical experiments often have distributions that are approximately normal. For example, a normal distribution will usually be a close approximation to the distribution of the heights or weights of individuals in a homogeneous population of people, corn stalks, or mice, or to the distribution of the tensile strength of pieces of steel produced by a certain process. Sometimes, a simple transformation of the observed random variables has a normal distribution.

The third reason for the preeminence of the normal distributions is the central limit theorem, which will be stated and proved in Sec. 6.3. If a large random sample is taken from some distribution, then even though this distribution is not itself approximately normal, a consequence of the central limit theorem is that many important functions of the observations in the sample will have distributions which are approximately normal. In particular, for a large random sample from any distribution that has a finite variance, the distribution of the average of the random sample will be approximately normal. We shall return to this topic in the next chapter.

Properties of Normal Distributions

Definition 5.6.1 Definition and p.d.f. A random variable X has the *normal distribution with mean* μ *and variance* σ^2 ($-\infty < \mu < \infty$ and $\sigma > 0$) if X has a continuous distribution with the following p.d.f.:

$$f(x|\mu, \sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad \text{for } -\infty < x < \infty. \tag{5.6.1}$$

We should first verify that the function defined in Eq. (5.6.1) is a p.d.f. Shortly thereafter, we shall verify that the mean and variance of the distribution with p.d.f. (5.6.1) are indeed μ and σ^2, respectively.

Theorem 5.6.1 The function defined in Eq. (5.6.1) is a p.d.f.

Proof Clearly, the function is nonnegative. We must also show that

$$\int_{-\infty}^{\infty} f(x|\mu, \sigma^2)\, dx = 1. \tag{5.6.2}$$

If we let $y = (x - \mu)/\sigma$, then

$$\int_{-\infty}^{\infty} f(x|\mu, \sigma^2)\, dx = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}y^2\right) dy.$$

We shall now let

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}y^2\right) dy. \tag{5.6.3}$$

Then we must show that $I = (2\pi)^{1/2}$.

From Eq. (5.6.3), it follows that

$$\begin{aligned} I^2 = I \cdot I &= \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}y^2\right) dy \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}z^2\right) dz \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(y^2 + z^2)\right] dy\, dz. \end{aligned}$$

We shall now change the variables in this integral from y and z to the polar coordinates r and θ by letting $y = r\cos\theta$ and $z = r\sin\theta$. Then, since $y^2 + z^2 = r^2$,

$$I^2 = \int_0^{2\pi} \int_0^{\infty} \exp\left(-\frac{1}{2}r^2\right) r\, dr\, d\theta = 2\pi, \tag{5.6.4}$$

where the inner integral in (5.6.4) is performed by substituting $v = r^2/2$ with $dv = r dr$, so the inner integral is

$$\int_0^{\infty} \exp(-v) dv = 1,$$

and the outer integral is 2π. Therefore, $I = (2\pi)^{1/2}$ and Eq. (5.6.2) has been established. ■

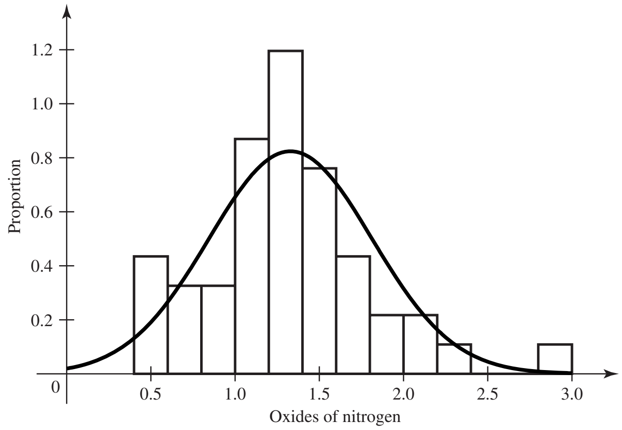
Example 5.6.2 Automobile Emissions. Consider the automobile engines described in Example 5.6.1. Figure 5.2 shows the histogram from Fig. 5.1 together with the normal p.d.f. having mean and variance chosen to match the observed data. Although the p.d.f. does not exactly match the shape of the histogram, it does correspond remarkably well. ◄

We could verify directly, using integration by parts, that the mean and variance of the distribution with p.d.f. given by Eq. (5.6.1) are, respectively, μ and σ^2. (See Exercise 26.) However, we need the moment generating function anyway, and then we can just take two derivatives of the m.g.f. to find the first two moments.

Theorem 5.6.2 Moment Generating Function. The m.g.f. of the distribution with p.d.f. given by Eq. (5.6.1) is

$$\psi(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \quad \text{for } -\infty < t < \infty. \tag{5.6.5}$$

Figure 5.2 Histogram of emissions of oxides of nitrogen for Example 5.6.2 together with a matching normal p.d.f.



Proof By the definition of an m.g.f.,

$$\psi(t) = E(e^{tX}) = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[tx - \frac{(x-\mu)^2}{2\sigma^2}\right] dx.$$

By completing the square inside the brackets (see Exercise 24), we obtain the relation

$$tx - \frac{(x-\mu)^2}{2\sigma^2} = \mu t + \frac{1}{2}\sigma^2 t^2 - \frac{[x-(\mu+\sigma^2 t)]^2}{2\sigma^2}.$$

Therefore,

$$\psi(t) = C \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right),$$

where

$$C = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{-\frac{[x-(\mu+\sigma^2 t)]^2}{2\sigma^2}\right\} dx.$$

If we now replace μ with $\mu + \sigma^2 t$ in Eq. (5.6.1), it follows from Eq. (5.6.2) that $C = 1$. Hence, the m.g.f. of the normal distribution is given by Eq. (5.6.5). ■

We are now ready to verify the mean and variance.

Theorem 5.6.3 Mean and Variance. The mean and variance of the distribution with p.d.f. given by Eq. (5.6.1) are μ and σ^2, respectively.

Proof The first two derivatives of the m.g.f. in Eq. (5.6.5) are

$$\psi'(t) = \left(\mu + \sigma^2 t\right) \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

$$\psi''(t) = \left([\mu + \sigma^2 t]^2 + \sigma^2\right) \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

Plugging $t = 0$ into each of these derivatives yields

$$E(X) = \psi'(0) = \mu \quad \text{and} \quad \mathrm{Var}(X) = \psi''(0) - [\psi'(0)]^2 = \sigma^2. \qquad ■$$

Since the m.g.f. $\psi(t)$ is finite for all values of t, all the moments $E(X^k)$ ($k = 1, 2, \ldots$) will also be finite.

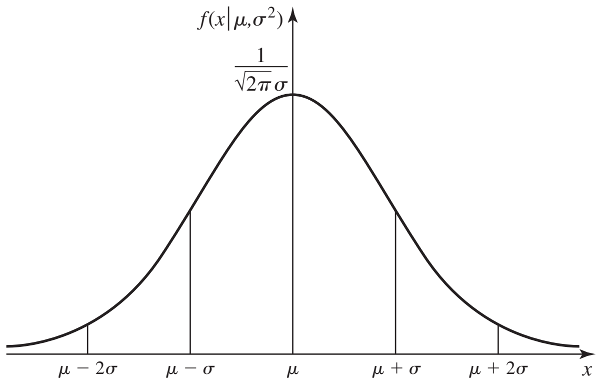


Figure 5.3 The p.d.f. of a normal distribution.

Example 5.6.3

Stock Price Changes. A popular model for the change in the price of a stock over a period of time of length u is to say that the price after time u is $S_u = S_0 e^{Z_u}$, where Z_u has the normal distribution with mean μu and variance $\sigma^2 u$. In this formula, S_0 is the present price of the stock, and σ is called the *volatility* of the stock price. The expected value of S_u can be computed from the m.g.f. ψ of Z_u:

$$E(S_u) = S_0 E(e^{Z_u}) = S_0 \psi(1) = S_0 e^{\mu u + \sigma^2 u/2}. \qquad ◀$$

The Shapes of Normal Distributions It can be seen from Eq. (5.6.1) that the p.d.f. $f(x|\mu, \sigma^2)$ of the normal distribution with mean μ and variance σ^2 is symmetric with respect to the point $x = \mu$. Therefore, μ is both the mean and the median of the distribution. Furthermore, μ is also the mode of the distribution. In other words, the p.d.f. $f(x|\mu, \sigma^2)$ attains its maximum value at the point $x = \mu$. Finally, by differentiating $f(x|\mu, \sigma^2)$ twice, it can be found that there are points of inflection at $x = \mu + \sigma$ and at $x = \mu - \sigma$.

The p.d.f. $f(x|\mu, \sigma^2)$ is sketched in Fig. 5.3. It is seen that the curve is "bell-shaped." However, it is not necessarily true that every arbitrary bell-shaped p.d.f. can be approximated by the p.d.f. of a normal distribution. For example, the p.d.f. of a Cauchy distribution, as sketched in Fig. 4.3, is a symmetric bell-shaped curve which apparently resembles the p.d.f. sketched in Fig. 5.3. However, since no moment of the Cauchy distribution—not even the mean—exists, the tails of the Cauchy p.d.f. must be quite different from the tails of the normal p.d.f.

Linear Transformations We shall now show that if a random variable X has a normal distribution, then every linear function of X will also have a normal distribution.

Theorem 5.6.4

If X has the normal distribution with mean μ and variance σ^2 and if $Y = aX + b$, where a and b are given constants and $a \neq 0$, then Y has the normal distribution with mean $a\mu + b$ and variance $a^2\sigma^2$.

Proof The m.g.f. ψ of X is given by Eq. (5.6.5). If ψ_Y denotes the m.g.f. of Y, then

$$\psi_Y(t) = e^{bt}\psi(at) = \exp\left[(a\mu + b)t + \frac{1}{2}a^2\sigma^2t^2\right] \quad \text{for } -\infty < t < \infty.$$

By comparing this expression for ψ_Y with the m.g.f. of a normal distribution given in Eq. (5.6.5), we see that ψ_Y is the m.g.f. of the normal distribution with mean $a\mu + b$ and variance $a^2\sigma^2$. Hence, Y must have this normal distribution. ■

The Standard Normal Distribution

Definition 5.6.2 Standard Normal Distribution. The normal distribution with mean 0 and variance 1 is called the *standard normal distribution*. The p.d.f. of the standard normal distribution is usually denoted by the symbol ϕ, and the c.d.f. is denoted by the symbol Φ. Thus,

$$\phi(x) = f(x|0, 1) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}x^2\right) \quad \text{for } -\infty < x < \infty \tag{5.6.6}$$

and

$$\Phi(x) = \int_{-\infty}^{x} \phi(u)\, du \quad \text{for } -\infty < x < \infty, \tag{5.6.7}$$

where the symbol u is used in Eq. (5.6.7) as a dummy variable of integration.

The c.d.f. $\Phi(x)$ cannot be expressed in closed form in terms of elementary functions. Therefore, probabilities for the standard normal distribution or any other normal distribution can be found only by numerical approximations or by using a table of values of $\Phi(x)$ such as the one given at the end of this book. In that table, the values of $\Phi(x)$ are given only for $x \ge 0$. Most computer packages that do statistical analysis contain functions that compute the c.d.f. and the quantile function of the standard normal distribution. Knowing the values of $\Phi(x)$ for $x \ge 0$ and $\Phi^{-1}(p)$ for $0.5 < p < 1$ is sufficient for calculating the c.d.f. and the quantile function of any normal distribution at any value, as the next two results show.

Theorem 5.6.5 Consequences of Symmetry. For all x and all $0 < p < 1$,

$$\Phi(-x) = 1 - \Phi(x) \quad \text{and} \quad \Phi^{-1}(p) = -\Phi^{-1}(1-p). \tag{5.6.8}$$

Proof Since the p.d.f. of the standard normal distribution is symmetric with respect to the point $x = 0$, it follows that $\Pr(X \le x) = \Pr(X \ge -x)$ for every number x $(-\infty < x < \infty)$. Since $\Pr(X \le x) = \Phi(x)$ and $\Pr(X \ge -x) = 1 - \Phi(-x)$, we have the first equation in Eq. (5.6.8). The second equation follows by letting $x = \Phi^{-1}(p)$ in the first equation and then applying the function Φ^{-1} to both sides of the equation. ■

Theorem 5.6.6 Converting Normal Distributions to Standard. Let X have the normal distribution with mean μ and variance σ^2. Let F be the c.d.f. of X. Then $Z = (X - \mu)/\sigma$ has the standard normal distribution, and, for all x and all $0 < p < 1$,

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \tag{5.6.9}$$

$$F^{-1}(p) = \mu + \sigma\Phi^{-1}(p). \tag{5.6.10}$$

Proof It follows immediately from Theorem 5.6.4 that $Z = (X - \mu)/\sigma$ has the standard normal distribution. Therefore,

$$F(x) = \Pr(X \le x) = \Pr\left(Z \le \frac{x-\mu}{\sigma}\right),$$

which establishes Eq. (5.6.9). For Eq. (5.6.10), let $p = F(x)$ in Eq. (5.6.9) and then solve for x in the resulting equation. ■

Example 5.6.4 Determining Probabilities for a Normal Distribution. Suppose that X has the normal distribution with mean 5 and standard deviation 2. We shall determine the value of $\Pr(1 < X < 8)$.

If we let $Z = (X - 5)/2$, then Z will have the standard normal distribution and

$$\Pr(1 < X < 8) = \Pr\left(\frac{1-5}{2} < \frac{X-5}{2} < \frac{8-5}{2}\right) = \Pr(-2 < Z < 1.5).$$

Furthermore,

$$\begin{aligned}\Pr(-2 < Z < 1.5) &= \Pr(Z < 1.5) - \Pr(Z \le -2)\\ &= \Phi(1.5) - \Phi(-2)\\ &= \Phi(1.5) - [1 - \Phi(2)].\end{aligned}$$

From the table at the end of this book, it is found that $\Phi(1.5) = 0.9332$ and $\Phi(2) = 0.9773$. Therefore,

$$\Pr(1 < X < 8) = 0.9105.$$ ◂

Example 5.6.5 Quantiles of Normal Distributions. Suppose that the engineers who collected the automobile emissions data in Example 5.6.1 are interested in finding out whether most engines are serious polluters. For example, they could compute the 0.05 quantile of the distribution of emissions and declare that 95 percent of the engines of the type tested exceed this quantile. Let X be the average grams of oxides of nitrogen per mile for a typical engine. Then the engineers modeled X as having a normal distribution. The normal distribution plotted in Fig. 5.2 has mean 1.329 and standard deviation 0.4844. The c.d.f. of X would then be $F(x) = \Phi([x - 1.329]/0.4844)$, and the quantile function would be $F^{-1}(p) = 1.329 + 0.4844\Phi^{-1}(p)$, where Φ^{-1} is the quantile function of the standard normal distribution, which can be evaluated using a computer or from tables. To find $\Phi^{-1}(p)$ from the table of Φ, find the closest value to p in the $\Phi(x)$ column and read the inverse from the x column. Since the table only has values of $p > 0.5$, we use Eq. (5.6.8) to conclude that $\Phi^{-1}(0.05) = -\Phi^{-1}(0.95)$. So, look up 0.95 in $\Phi(x)$ column (halfway between 0.9495 and 0.9505) to find $x = 1.645$ (halfway between 1.64 and 1.65) and conclude that $\Phi^{-1}(0.05) = -1.645$. The 0.05 quantile of X is then $1.329 + 0.4844 \times (-1.645) = 0.5322$. ◂

Comparisons of Normal Distributions

The p.d.f.'s of three normal distributions are sketched in Fig. 5.4 for a fixed value of μ and three different values of σ ($\sigma = 1/2$, 1, and 2). It can be seen from this figure that the p.d.f. of a normal distribution with a small value of σ has a high peak and is very concentrated around the mean μ, whereas the p.d.f. of a normal distribution with a larger value of σ is relatively flat and is spread out more widely over the real line.

An important fact is that every normal distribution contains the same total amount of probability within one standard deviation of its mean, the same amount within two standard deviations of its mean, and the same amount within any other fixed number of standard deviations of its mean. In general, if X has the normal distribution with mean μ and variance σ^2, and if Z has the standard normal distribution, then for $k > 0$,

$$p_k = \Pr(|X - \mu| \le k\sigma) = \Pr(|Z| \le k).$$

In Table 5.2, the values of this probability p_k are given for various values of k. These probabilities can be computed from a table of Φ or using computer programs.

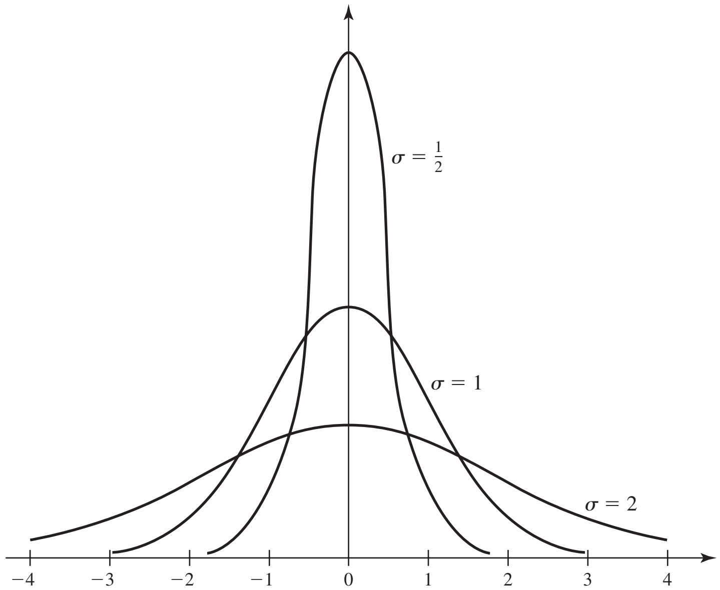


Figure 5.4 The normal p.d.f. for $\mu = 0$ and $\sigma = 1/2, 1, 2$.

Table 5.2 Probabilities that normal random variables are within k standard deviations of their means

k	p_k
1	0.6826
2	0.9544
3	0.9974
4	0.99994
5	$1 - 6 \times 10^{-7}$
10	$1 - 2 \times 10^{-23}$

Although the p.d.f. of a normal distribution is positive over the entire real line, it can be seen from this table that the total amount of probability outside an interval of four standard deviations on each side of the mean is only 0.00006.

Linear Combinations of Normally Distributed Variables

In the next theorem and corollary, we shall prove the following important result: Every linear combination of random variables that are independent and normally distributed will also have a normal distribution.

Theorem 5.6.7 If the random variables $X_1, \dots, X_k$ are independent and if X_i has the normal distribution with mean μ_i and variance σ_i^2 $(i = 1, \dots, k)$, then the sum $X_1 + \cdots + X_k$ has the normal distribution with mean $\mu_1 + \cdots + \mu_k$ and variance $\sigma_1^2 + \cdots + \sigma_k^2$.

Proof Let $\psi_i(t)$ denote the m.g.f. of X_i for $i = 1, \dots, k$, and let $\psi(t)$ denote the m.g.f. of $X_1 + \cdots + X_k$. Since the variables $X_1, \dots, X_k$ are independent, then

$$\begin{aligned}\psi(t) &= \prod_{i=1}^{k} \psi_i(t) = \prod_{i=1}^{k} \exp\left(\mu_i t + \frac{1}{2}\sigma_i^2 t^2\right) \\ &= \exp\left[\left(\sum_{i=1}^{k} \mu_i\right)t + \frac{1}{2}\left(\sum_{i=1}^{k} \sigma_i^2\right)t^2\right] \quad \text{for } -\infty < t < \infty.\end{aligned}$$

From Eq. (5.6.5), the m.g.f. $\psi(t)$ can be identified as the m.g.f. of the normal distribution for which the mean is $\sum_{i=1}^{k} \mu_i$ and the variance is $\sum_{i=1}^{k} \sigma_i^2$. Hence, the distribution of $X_1 + \cdots + X_k$ must be as stated in the theorem. ■

The following corollary is now obtained by combining Theorems 5.6.4 and 5.6.7.

Corollary 5.6.1 If the random variables $X_1, \dots, X_k$ are independent, if X_i has the normal distribution with mean μ_i and variance σ_i^2 $(i = 1, \dots, k)$, and if $a_1, \dots, a_k$ and b are constants for which at least one of the values $a_1, \dots, a_k$ is different from 0, then the variable $a_1X_1 + \cdots + a_kX_k + b$ has the normal distribution with mean $a_1\mu_1 + \cdots + a_k\mu_k + b$ and variance $a_1^2\sigma_1^2 + \cdots + a_k^2\sigma_k^2$. ■

Example 5.6.6 Heights of Men and Women. Suppose that the heights, in inches, of the women in a certain population follow the normal distribution with mean 65 and standard deviation 1, and that the heights of the men follow the normal distribution with mean 68 and standard deviation 3. Suppose also that one woman is selected at random and, independently, one man is selected at random. We shall determine the probability that the woman will be taller than the man.

Let W denote the height of the selected woman, and let M denote the height of the selected man. Then the difference $W - M$ has the normal distribution with mean $65 - 68 = -3$ and variance $1^2 + 3^2 = 10$. Therefore, if we let

$$Z = \frac{1}{10^{1/2}}(W - M + 3),$$

then Z has the standard normal distribution. It follows that

$$\begin{aligned}\Pr(W > M) &= \Pr(W - M > 0) \\ &= \Pr\left(Z > \frac{3}{10^{1/2}}\right) = \Pr(Z > 0.949) \\ &= 1 - \Phi(0.949) = 0.171.\end{aligned}$$

Thus, the probability that the woman will be taller than the man is 0.171. ◀

Averages of random samples of normal random variables figure prominently in many statistical calculations. To fix notation, we start with a general defintion.

Definition 5.6.3 Sample Mean. Let $X_1, \dots, X_n$ be random variables. The average of these n random variables, $\frac{1}{n}\sum_{i=1}^{n} X_i$, is called their *sample mean* and is commonly denoted $\overline{X}_n$.

The following simple corollary to Corollary 5.6.1 gives the distribution of the sample mean of a random sample of normal random variables.

Corollary 5.6.2 Suppose that the random variables $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2, and let $\overline{X}_n$ denote their sample mean. Then $\overline{X}_n$ has the normal distribution with mean μ and variance σ^2/n.

Proof Since $\overline{X}_n = \sum_{i=1}^{n}(1/n)X_i$, it follows from Corollary 5.6.1 that the distribution of $\overline{X}_n$ is normal with mean $\sum_{i=1}^{n}(1/n)\mu = \mu$ and variance $\sum_{i=1}^{n}(1/n)^2\sigma^2 = \sigma^2/n$. ■

Example 5.6.7 Determining a Sample Size. Suppose that a random sample of size n is to be taken from the normal distribution with mean μ and variance 9. (The heights of men in Example 5.6.6 have such a distribution with $\mu = 68$.) We shall determine the minimum value of n for which

$$\Pr(|\overline{X}_n - \mu| \le 1) \ge 0.95.$$

It is known from Corollary 5.6.2 that the sample mean $\overline{X}_n$ will have the normal distribution for which the mean is μ and the standard deviation is $3/n^{1/2}$. Therefore, if we let

$$Z = \frac{n^{1/2}}{3}(\overline{X}_n - \mu),$$

then Z will have the standard normal distribution. In this example, n must be chosen so that

$$\Pr(|\overline{X}_n - \mu| \le 1) = \Pr\left(|Z| \le \frac{n^{1/2}}{3}\right) \ge 0.95. \tag{5.6.11}$$

For each positive number x, it will be true that $\Pr(|Z| \le x) \ge 0.95$ if and only if $1 - \Phi(x) = \Pr(Z > x) \le 0.025$. From the table of the standard normal distribution at the end of this book, it is found that $1 - \Phi(x) \le 0.025$ if and only if $x \ge 1.96$. Therefore, the inequality in relation (5.6.11) will be satisfied if and only if

$$\frac{n^{1/2}}{3} \ge 1.96.$$

Since the smallest permissible value of n is 34.6, the sample size must be at least 35 in order that the specified relation will be satisfied. ◄

Example 5.6.8 Interval for Mean. Consider a popluation with a normal distribution such as the heights of men in Example 5.6.6. Suppose that we are not willing to specify the precise distribution as we did in that example, but rather only that the standard deviation is 3, leaving the mean μ unspecified. If we sample a number of men from this population, we could try to use their sampled heights to give us some idea what μ equals. A popular form of statistical inference that will be discussed in Sec. 8.5 finds an interval that has a specified probability of containing μ. To be specific, suppose that we observe a random sample of size n from the normal distribution with mean μ and standard deviation 3. Then, $\overline{X}_n$ has the normal distribution with mean μ and standard deviation $3/n^{1/2}$ as in Example 5.6.7. Similarly, we can define

$$Z = \frac{n^{1/2}}{3}(\overline{X}_n - \mu),$$

which then has the standard normal distribution. Hence,

$$0.95 = \Pr(|Z| < 1.96) = \Pr\left(|\overline{X}_n - \mu| < 1.96\frac{3}{n^{1/2}}\right). \tag{5.6.12}$$

It is easy to verify that

$$|\overline{X}_n - \mu| < 1.96\frac{3}{n^{1/2}} \text{ if and only if}$$

$$\overline{X}_n - 1.96\frac{3}{n^{1/2}} < \mu < \overline{X}_n + 1.96\frac{3}{n^{1/2}}. \tag{5.6.13}$$

The two inequalities in Eq. (5.6.13) hold if and only if the interval

$$\left(\overline{X}_n - 1.96\frac{3}{n^{1/2}}, \overline{X}_n + 1.96\frac{3}{n^{1/2}}\right) \tag{5.6.14}$$

contains the value of μ. It follows from Eq. (5.6.12) that the probability is 0.95 that the interval in (5.6.14) contains μ. Now, suppose that the sample size is $n = 36$. Then the half-width of the interval (5.6.14) is then $3/36^{1/2} = 0.98$. We will not know the endpoints of the interval until after we observe $\overline{X}_n$. However, we know now that the interval $\left(\overline{X}_n - 0.98, \overline{X}_n + 0.98\right)$ has probability 0.95 of containing μ. ◄

The Lognormal Distributions

It is very common to use normal distributions to model logarithms of random variables. For this reason, a name is given to the distribution of the original random variables before transforming.

Definition 5.6.4 Lognormal Distribution. If $\log(X)$ has the normal distribution with mean μ and variance σ^2, we say that X has the *lognormal distribution* with parameters μ and σ^2.

Example 5.6.9 Failure Times of Ball Bearings. Products that are subject to wear and tear are generally tested for endurance in order to estimate their useful lifetimes. Lawless (1982, example 5.2.2) describes data taken from Lieblein and Zelen (1956), which are measurements of the numbers of millions of revolutions before failure for 23 ball bearings. The lognormal distribution is one popular model for times until failure. Figure 5.5 shows a histogram of the 23 lifetimes together with a lognormal p.d.f. with parameters chosen to match the observed data. The bars of the histogram in Fig. 5.5 have areas that equal the proportions of the sample that lie between the points on the horizontal axis where the sides of the bars stand. Suppose that the engineers are interested in knowing how long to wait until there is a 90 percent chance that a ball

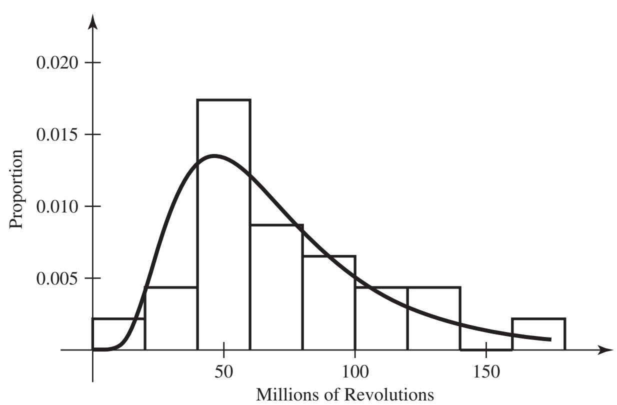


Figure 5.5 Histogram of lifetimes of ball bearings and fitted lognormal p.d.f. for Example 5.6.9.

bearing will have failed. Then they want the 0.9 quantile of the distribution of lifetimes. Let X be the time to failure of a ball bearing. The lognormal distribution of X plotted in Fig. 5.5 has parameters 4.15 and 0.5334^2. The c.d.f. of X would then be $F(x) = \Phi([\log(x) - 4.15]/0.5334)$, and the quantile function would be

$$F^{-1}(p) = e^{4.15+0.5334\Phi^{-1}(p)},$$

where Φ^{-1} is the quantile function of the standard normal distribution. With $p = 0.9$, we get $\Phi^{-1}(0.9) = 1.28$ and $F^{-1}(0.9) = 125.6$. ◄

The moments of a lognormal random variable are easy to compute based on the m.g.f. of a normal distribution. If $Y = \log(X)$ has the normal distribution with mean μ and variance σ^2, then the m.g.f. of Y is $\psi(t) = \exp(\mu t + 0.5\sigma^2 t^2)$. However, the definition of ψ is $\psi(t) = E(e^{tY})$. Since $Y = \log(X)$, we have

$$\psi(t) = E(e^{tY}) = E(e^{t\log(X)}) = E(X^t).$$

It follows that $E(X^t) = \psi(t)$ for all real t. In particular, the mean and variance of X are

$$E(X) = \psi(1) = \exp(\mu + 0.5\sigma^2), \tag{5.6.15}$$
$$\mathrm{Var}(X) = \psi(2) - \psi(1)^2 = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1].$$

Example 5.6.10 **Stock and Option Prices.** Consider a stock like the one in Example 5.6.3 whose current price is S_0. Suppose that the price at u time units in the future is $S_u = S_0 e^{Z_u}$, where Z_u has the normal distribution with mean μu and variance $\sigma^2 u$. Note that $S_0 e^{Z_u} = e^{Z_u + \log(S_0)}$ and $Z_u + \log(S_0)$ has the normal distribution with mean $\mu u + \log(S_0)$ and variance $\sigma^2 u$. So S_u has the lognormal distribution with parameters $\mu u + \log(S_0)$ and $\sigma^2 u$.

Black and Scholes (1973) developed a pricing scheme for options on stocks whose prices follow a lognormal distribution. For the remainder of this example, we shall consider a single time u and write the stock price as $S_u = S_0 e^{\mu u + \sigma u^{1/2} Z}$, where Z has the standard normal distribution. Suppose that we need to price the option to buy one share of the above stock for the price q at a particular time u in the future. As in Example 4.1.14 on page 214, we shall use risk-neutral pricing. That is, we force the present value of $E(S_u)$ to equal S_0. If u is measured in years and the risk-free interest rate is r per year, then the present value of $E(S_u)$ is $e^{-ru}E(S_u)$. (This assumes that compounding of interest is done continuously instead of just once as it was in Example 4.1.14. The effect of continuous compounding is examined in Exercise 25.) But $E(S_u) = S_0 e^{\mu u + \sigma^2 u/2}$. Setting S_0 equal to $e^{-ru} S_0 e^{\mu u + \sigma^2 u/2}$ yields $\mu = r - \sigma^2/2$ when doing risk-neutral pricing.

Now we can determine a price for the specified option. The value of the option at time u will be $h(S_u)$, where

$$h(s) = \begin{cases} s - q & \text{if } s > q, \\ 0 & \text{otherwise.} \end{cases}$$

Set $\mu = r - \sigma^2/2$, and it is easy to see that $h(S_u) > 0$ if and only if

$$Z > \frac{\log\left(\frac{q}{S_0}\right) - (r - \sigma^2/2)u}{\sigma u^{1/2}}. \tag{5.6.16}$$

We shall refer to the constant on the right-hand side of Eq. (5.6.16) as c. The risk-neutral price of the option is the present value of $E(h(S_u))$, which equals

$$e^{-ru}E[h(S_u)] = e^{-ru}\int_c^\infty \left[S_0 e^{[r-\sigma^2/2]u+\sigma u^{1/2}z} - q\right]\frac{1}{(2\pi)^{1/2}}e^{-z^2/2}\,dz. \tag{5.6.17}$$

To compute the integral in Eq. (5.6.17), split the integrand into two parts at the $-q$. The second integral is then just a constant times the integral of a normal p.d.f., namely,

$$-e^{-ru}q\int_c^\infty \frac{1}{(2\pi)^{1/2}}e^{-z^2}\,dz = -e^{-ru}q[1-\Phi(c)].$$

The first integral in Eq. (5.6.17), is

$$e^{-\sigma^2u/2}S_0\int_c^\infty \frac{1}{(2\pi)^{1/2}}e^{-z^2/2+\sigma u^{1/2}z}\,dz.$$

This can be converted into the integral of a normal p.d.f. times a constant by completing the square (see Exercise 24). The result of completing the square is

$$e^{-\sigma^2u/2}S_0\int_c^\infty \frac{1}{(2\pi)^{1/2}}e^{-(z-\sigma u^{1/2})^2/2+\sigma^2u/2}\,dz = S_0[1-\Phi(c-\sigma u^{1/2})].$$

Finally, combine the two integrals into the option price, using the fact that $1-\Phi(x) = \Phi(-x)$:

$$S_0\Phi(\sigma u^{1/2}-c) - qe^{-ru}\Phi(-c). \tag{5.6.18}$$

This is the famous *Black-Scholes formula* for pricing options. As a simple example, suppose that $q = S_0$, $r = 0.06$ (6 percent interest), $u = 1$ (one year wait), and $\sigma = 0.1$. Then (5.6.18) says that the option price should be $0.0746S_0$. If the distribution of S_u is different from the form used here, simulation techniques (see Chapter 12) can be used to help price options. ◀

The p.d.f.'s of the lognormal distributions will be found in Exercise 17 of this section. The c.d.f. of each lognormal distribution is easily constructed from the standard normal c.d.f. Φ. Let X have the lognormal distribution with parameters μ and σ^2. Then

$$\Pr(X \le x) = \Pr(\log(X) \le \log(x)) = \Phi\left(\frac{\log(x)-\mu}{\sigma}\right).$$

The results from earlier in this section about linear combinations of normal random variables translate into results about products of powers of lognormal random variables. Results about sums of independent normal random variables translate into results about products of independent lognormal random variables.

Summary

We introduced the family of normal distributions. The parameters of each normal distribution are its mean and variance. A linear combination of independent normal random variables has the normal distribution with mean equal to the linear combination of the means and variance determined by Corollary 4.3.1. In particular, if X has the normal distribution with mean μ and variance σ^2, then $(X-\mu)/\sigma$ has the standard normal distribution (mean 0 and variance 1). Probabilities and quantiles for normal distributions can be obtained from tables or computer programs for standard normal probabilities and quantiles. For example, if X has the normal distribution with mean μ and variance σ^2, then the c.d.f. of X is $F(x) = \Phi([x-\mu]/\sigma)$ and the quantile function of X is $F^{-1}(p) = \mu + \Phi^{-1}(p)\sigma$, where Φ is the standard normal c.d.f.

Exercises

1. Find the 0.5, 0.25, 0.75, 0.1, and 0.9 quantiles of the standard normal distribution.

2. Suppose that X has the normal distribution for which the mean is 1 and the variance is 4. Find the value of each of the following probabilities:

a. $\Pr(X \le 3)$ **b.** $\Pr(X > 1.5)$
c. $\Pr(X = 1)$ **d.** $\Pr(2 < X < 5)$
e. $\Pr(X \ge 0)$ **f.** $\Pr(-1 < X < 0.5)$
g. $\Pr(|X| \le 2)$ **h.** $\Pr(1 \le -2X + 3 \le 8)$

3. If the temperature in degrees Fahrenheit at a certain location is normally distributed with a mean of 68 degrees and a standard deviation of 4 degrees, what is the distribution of the temperature in degrees Celsius at the same location?

4. Find the 0.25 and 0.75 quantiles of the Fahrenheit temperature at the location mentioned in Exercise 3.

5. Let X_1, X_2, and X_3 be independent lifetimes of memory chips. Suppose that each X_i has the normal distribution with mean 300 hours and standard deviation 10 hours. Compute the probability that at least one of the three chips lasts at least 290 hours.

6. If the m.g.f. of a random variable X is $\psi(t) = e^{t^2}$ for $-\infty < t < \infty$, what is the distribution of X?

7. Suppose that the measured voltage in a certain electric circuit has the normal distribution with mean 120 and standard deviation 2. If three independent measurements of the voltage are made, what is the probability that all three measurements will lie between 116 and 118?

8. Evaluate the integral $\int_0^\infty e^{-3x^2}\,dx$.

9. A straight rod is formed by connecting three sections A, B, and C, each of which is manufactured on a different machine. The length of section A, in inches, has the normal distribution with mean 20 and variance 0.04. The length of section B, in inches, has the normal distribution with mean 14 and variance 0.01. The length of section C, in inches, has the normal distribution with mean 26 and variance 0.04. As indicated in Fig. 5.6, the three sections are joined so that there is an overlap of 2 inches at each connection. Suppose that the rod can be used in the construction of an airplane wing if its total length in inches is between 55.7 and 56.3. What is the probability that the rod can be used?

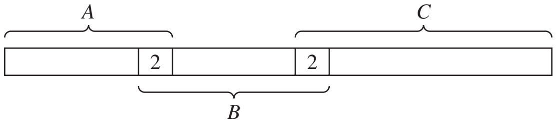


Figure 5.6 Sections of the rod in Exercise 9.

10. If a random sample of 25 observations is taken from the normal distribution with mean μ and standard deviation 2, what is the probability that the sample mean will lie within one unit of μ?

11. Suppose that a random sample of size n is to be taken from the normal distribution with mean μ and standard deviation 2. Determine the smallest value of n such that

$$\Pr(|\overline{X}_n - \mu| < 0.1) \ge 0.9.$$

12.

a. Sketch the c.d.f. Φ of the standard normal distribution from the values given in the table at the end of this book.

b. From the sketch given in part (a) of this exercise, sketch the c.d.f. of the normal distribution for which the mean is -2 and the standard deviation is 3.

13. Suppose that the diameters of the bolts in a large box follow a normal distribution with a mean of 2 centimeters and a standard deviation of 0.03 centimeter. Also, suppose that the diameters of the holes in the nuts in another large box follow the normal distribution with a mean of 2.02 centimeters and a standard deviation of 0.04 centimeter. A bolt and a nut will fit together if the diameter of the hole in the nut is greater than the diameter of the bolt and the difference between these diameters is not greater than 0.05 centimeter. If a bolt and a nut are selected at random, what is the probability that they will fit together?

14. Suppose that on a certain examination in advanced mathematics, students from university A achieve scores that are normally distributed with a mean of 625 and a variance of 100, and students from university B achieve scores which are normally distributed with a mean of 600 and a variance of 150. If two students from university A and three students from university B take this examination, what is the probability that the average of the scores of the two students from university A will be greater than the average of the scores of the three students from university B? *Hint:* Determine the distribution of the difference between the two averages.

15. Suppose that 10 percent of the people in a certain population have the eye disease glaucoma. For persons who have glaucoma, measurements of eye pressure X will be normally distributed with a mean of 25 and a variance of 1. For persons who do not have glaucoma, the pressure X will be normally distributed with a mean of 20 and a variance of 1. Suppose that a person is selected at random from the population and her eye pressure X is measured.

a. Determine the conditional probability that the person has glaucoma given that $X = x$.

b. For what values of x is the conditional probability in part (a) greater than 1/2?

16. Suppose that the joint p.d.f. of two random variables X and Y is

$$f(x, y) = \frac{1}{2\pi} e^{-(1/2)(x^2+y^2)} \quad \text{for } -\infty < x < \infty \text{ and } -\infty < y < \infty.$$

Find $\Pr(-\sqrt{2} < X + Y < 2\sqrt{2})$.

17. Consider a random variable X having the lognormal distribution with parameters μ and σ^2. Determine the p.d.f. of X.

18. Suppose that the random variables X and Y are independent and that each has the standard normal distribution. Show that the quotient X/Y has the Cauchy distribution.

19. Suppose that the measurement X of pressure made by a device in a particular system has the normal distribution with mean μ and variance 1, where μ is the true pressure. Suppose that the true pressure μ is unknown but has the uniform distribution on the interval [5, 15]. If $X = 8$ is observed, find the conditional p.d.f. of μ given $X = 8$.

20. Let X have the lognormal distribution with parameters 3 and 1.44. Find the probability that $X \le 6.05$.

21. Let X and Y be independent random variables such that $\log(X)$ has the normal distribution with mean 1.6 and variance 4.5 and $\log(Y)$ has the normal distribution with mean 3 and variance 6. Find the distribution of the product XY.

22. Suppose that X has the lognormal distribution with parameters μ and σ^2. Find the distribution of $1/X$.

23. Suppose that X has the lognormal distribution with parameters 4.1 and 8. Find the distribution of $3X^{1/2}$.

24. The method of *completing the square* is used several times in this text. It is a useful method for combining several quadratic and linear polynomials into a perfect square plus a constant. Prove the following identity, which is one general form of completing the square:

$$\begin{aligned}
\sum_{i=1}^{n} a_i(x - b_i)^2 + cx
&= \left(\sum_{i=1}^{n} a_i\right)\left(x - \frac{\sum_{i=1}^{n} a_i b_i - c/2}{\sum_{i=1}^{n} a_i}\right)^2 \\
&\quad + \sum_{i=1}^{n} a_i\left(b_i - \frac{\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} a_i}\right)^2 \\
&\quad + \left(\sum_{i=1}^{n} a_i\right)^{-1}\left[c\sum_{i=1}^{n} a_i b_i - c^2/4\right]
\end{aligned}$$

if $\sum_{i=1}^{n} a_i \ne 0$.

25. In Example 5.6.10, we considered the effect of continuous compounding of interest. Suppose that S_0 dollars earn a rate of r per year componded continuously for u years. Prove that the principal plus interest at the end of this time equals $S_0 e^{ru}$. *Hint:* Suppose that interest is compounded n times at intervals of u/n years each. At the end of each of the n intervals, the principal gets multiplied by $1 + ru/n$. Take the limit of the result as $n \to \infty$.

26. Let X have the normal distribution whose p.d.f. is given by (5.6.6). Instead of using the m.g.f., derive the variance of X using integration by parts.

5.7 The Gamma Distributions

The family of gamma distributions is a popular model for random variables that are known to be positive. The family of exponential distributions is a subfamily of the gamma distributions. The times between successive occurrences in a Poisson process have an exponential distribution. The gamma function, related to the gamma distributions, is an extension of factorials from integers to all positive numbers.

The Gamma Function

Example 5.7.1

Mean and Variance of Lifetime of a Light Bulb. Suppopse that we model the lifetime of a light bulb as a continuous random variable with the following p.d.f.:

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

If we wish to compute the mean and variance of such a lifetime, we need to compute the following integrals:

$$\int_0^\infty xe^{-x}dx, \quad \text{and} \quad \int_0^\infty x^2e^{-x}dx. \tag{5.7.1}$$

These integrals are special cases of an important function that we examine next. ◂

Definition 5.7.1 **The Gamma Function.** For each positive number α, let the value $\Gamma(\alpha)$ be defined by the following integral:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx. \tag{5.7.2}$$

The function Γ defined by Eq. (5.7.2) for $\alpha > 0$ is called the *gamma function*.

As an example,

$$\Gamma(1) = \int_0^\infty e^{-x}\,dx = 1. \tag{5.7.3}$$

The following result, together with Eq. (5.7.3), shows that $\Gamma(\alpha)$ is finite for every value of $\alpha > 0$.

Theorem 5.7.1 If $\alpha > 1$, then

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1). \tag{5.7.4}$$

Proof We shall apply the method of integration by parts to the integral in Eq. (5.7.2). If we let $u = x^{\alpha-1}$ and $dv = e^{-x}\,dx$, then $du = (\alpha - 1)x^{\alpha-2}\,dx$ and $v = -e^{-x}$. Therefore,

$$\begin{aligned}\Gamma(\alpha) &= \int_0^\infty u\,dv = [uv]_0^\infty - \int_0^\infty v\,du \\ &= [-x^{\alpha-1}e^{-x}]_{x=0}^\infty + (\alpha - 1)\int_0^\infty x^{\alpha-2}e^{-x}\,dx \\ &= 0 + (\alpha - 1)\Gamma(\alpha - 1).\end{aligned}$$

■

For integer values of α, we have a simple expression for the gamma function.

Theorem 5.7.2 For every positive integer n,

$$\Gamma(n) = (n - 1)!. \tag{5.7.5}$$

Proof It follows from Theorem 5.7.1 that for every integer $n \ge 2$,

$$\begin{aligned}\Gamma(n) &= (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) \\ &= (n-1)(n-2)\cdots 1\cdot\Gamma(1) \\ &= (n-1)!\Gamma(1).\end{aligned}$$

Since $\Gamma(1) = 1 = 0!$ by Eq. (5.7.3), the proof is complete. ■

Example 5.7.2 **Mean and Variance of Lifetime of a Light Bulb.** The two integrals in (5.7.1) are, respectively, $\Gamma(2) = 1! = 1$ and $\Gamma(3) = 2! = 2$. It follows that the mean of each lifetime is 1, and the variance is $2 - 1^2 = 1$. ◂

In many statistical applications, $\Gamma(\alpha)$ must be evaluated when α is either a positive integer or of the form $\alpha = n + (1/2)$ for some positive integer n. It follows from

Eq. (5.7.4) that for each positive integer n,

$$\Gamma\left(n+\frac{1}{2}\right)=\left(n-\frac{1}{2}\right)\left(n-\frac{3}{2}\right)\cdots\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right). \tag{5.7.6}$$

Hence, it will be possible to determine the value of $\Gamma\left(n+\frac{1}{2}\right)$ if we can evaluate $\Gamma\left(\frac{1}{2}\right)$.

From Eq. (5.7.2),

$$\Gamma\left(\frac{1}{2}\right)=\int_0^\infty x^{-1/2}e^{-x}\,dx.$$

If we let $x=(1/2)y^2$ in this integral, then $dx=y\,dy$ and

$$\Gamma\left(\frac{1}{2}\right)=2^{1/2}\int_0^\infty \exp\left(-\frac{1}{2}y^2\right)dy. \tag{5.7.7}$$

Because the integral of the p.d.f. of the standard normal distribution is equal to 1, it follows that

$$\int_{-\infty}^\infty \exp\left(-\frac{1}{2}y^2\right)dy=(2\pi)^{1/2}. \tag{5.7.8}$$

Because the integrand in (5.7.8) is symmetric around $y=0$,

$$\int_0^\infty \exp\left(-\frac{1}{2}y^2\right)dy=\frac{1}{2}(2\pi)^{1/2}=\left(\frac{\pi}{2}\right)^{1/2}.$$

It now follows from Eq. (5.7.7) that

$$\Gamma\left(\frac{1}{2}\right)=\pi^{1/2}. \tag{5.7.9}$$

For example, it is found from Eqs. (5.7.6) and (5.7.9) that

$$\Gamma\left(\frac{7}{2}\right)=\left(\frac{5}{2}\right)\left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\pi^{1/2}=\frac{15}{8}\pi^{1/2}.$$

We present two final useful results before we introduce the gamma distributions.

Theorem 5.7.3 For each $\alpha>0$ and each $\beta>0$,

$$\int_0^\infty x^{\alpha-1}\exp(\beta x)dx=\frac{\Gamma(\alpha)}{\beta^\alpha}. \tag{5.7.10}$$

Proof Make the change of variables $y=\beta x$ so that $x=y/\beta$ and $dx=dy/\beta$. The result now follows easily from Eq. (5.7.2). ■

There is a version of Stirling's formula (Theorem 1.7.5) for the gamma function, which we state without proof.

Theorem 5.7.4 Stirling's Formula. $\lim_{x\to\infty}\frac{(2\pi)^{1/2}x^{x-1/2}e^{-x}}{\Gamma(x)}=1.$ ■

Example 5.7.3 Service Times in a Queue. For $i=1,\ldots,n$, suppose that customer i in a queue must wait time X_i for service once reaching the head of the queue. Let Z be the rate at which the average customer is served. A typical probability model for this situation

is to say that, conditional on $Z = z$, $X_1, \ldots, X_n$ are i.i.d. with a distribution having the conditional p.d.f. $g_1(x_i|z) = z\exp(-zx_i)$ for $x_i > 0$. Suppose that Z is also unknown and has the p.d.f. $f_2(z) = 2\exp(-2z)$ for $z > 0$. The joint p.d.f. of $X_1, \ldots, X_n, Z$ is then

$$\begin{aligned} f(x_1, \ldots, x_n, z) &= \prod_{i=1}^{n} g_1(x_i|z) f_2(z) \\ &= 2z^n \exp\left(-z\left[2 + x_1 + \cdots + x_n\right]\right), \end{aligned} \tag{5.7.11}$$

if $z, x_1, \ldots, x_n > 0$ and 0 otherwise. In order to calculate the marginal joint distribution of $X_1, \ldots, X_n$, we must integrate z out of the the joint p.d.f. above. We can apply Theorem 5.7.3 with $\alpha = n + 1$ and $\beta = 2 + x_1 + \cdots + x_n$ together with Theorem 5.7.2 to integrate the function in Eq. (5.7.11). The result is

$$\int_0^\infty f(x_1, \ldots, x_n, z)dz = \frac{2(n!)}{\left(2 + \sum_{i=1}^{n} x_i\right)^{n+1}}, \tag{5.7.12}$$

for all $x_i > 0$ and 0 otherwise. This is the same joint p.d.f. that was used in Example 3.7.5 on page 154. ◀

The Gamma Distributions

Example 5.7.4 **Service Times in a Queue.** In Example 5.7.3, suppose that we observe the service times of n customers and want to find the conditional distribution of the rate Z. We can easily find the conditional p.d.f. $g_2(z|x_1, \ldots, x_n)$ of Z given $X_1 = x_1, \ldots, X_n = x_n$ by dividing the joint p.d.f. of $X_1, \ldots, X_n, Z$ in Eq. (5.7.11) by the p.d.f. of $X_1, \ldots, X_n$ in Eq. (5.7.12). The calculation is simplified by defining $y = 2 + \sum_{i=1}^{n} x_i$. We then obtain

$$g_2(z|x_1, \ldots, x_n) = \begin{cases} \dfrac{y^{n+1}}{n!} e^{-yz}, & \text{for } z > 0, \\ 0 & \text{otherwise.} \end{cases}$$ ◀

Distributions with p.d.f.'s like the one at the end of Example 5.7.4 are members of a commonly used family, which we now define.

Definition 5.7.2 **Gamma Distributions.** Let α and β be positive numbers. A random variable X has the *gamma distribution with parameters α and β* if X has a continuous distribution for which the p.d.f. is

$$f(x|\alpha, \beta) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{for } x > 0, \\ 0 & \text{for } x \le 0. \end{cases} \tag{5.7.13}$$

That the integral of the p.d.f. in Eq. (5.7.13) is 1 follows easily from Theorem 5.7.3.

Example 5.7.5 **Service Times in a Queue.** In Example 5.7.4, we can easily recognize the conditional p.d.f. as the p.d.f. of the gamma distribution with parameters $\alpha = n + 1$ and $\beta = y$. ◀

If X has a gamma distribution, then the moments of X are easily found from Eqs. (5.7.13) and (5.7.10).

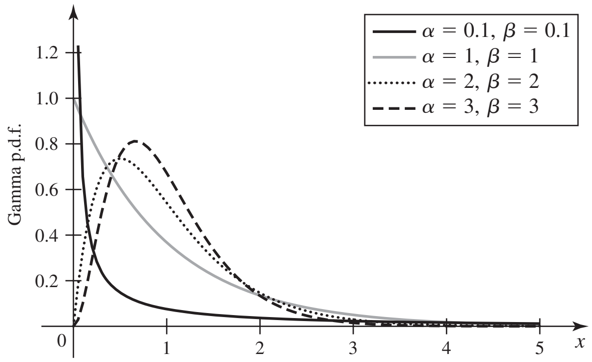


Figure 5.7 Graphs of the p.d.f.'s of several different gamma distributions with common mean of 1.

Theorem 5.7.5 Moments. Let X have the gamma distribution with parameters α and β. For $k = 1, 2, \ldots,$

$$E(X^k) = \frac{\Gamma(\alpha+k)}{\beta^k\Gamma(\alpha)} = \frac{\alpha(\alpha+1)\cdots(\alpha+k-1)}{\beta^k}.$$

In particular, $E(X) = \frac{\alpha}{\beta}$, and $\text{Var}(X) = \frac{\alpha}{\beta^2}$.

Proof For $k = 1, 2, \ldots,$

$$E(X^k) = \int_0^\infty x^k f(x|\alpha,\beta)\,dx = \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha+k-1}e^{-\beta x}\,dx$$
$$= \frac{\beta^\alpha}{\Gamma(\alpha)}\cdot\frac{\Gamma(\alpha+k)}{\beta^{\alpha+k}} = \frac{\Gamma(\alpha+k)}{\beta^k\Gamma(\alpha)}. \tag{5.7.14}$$

The expression for $E(X)$ follows immediately from (5.7.14). The variance can be computed as

$$\text{Var}(X) = \frac{\alpha(\alpha+1)}{\beta^2} - \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2}.$$ ■

Figure 5.7 shows several gamma distribution p.d.f.'s that all have mean equal to 1 but different values of α and β.

Example 5.7.6 Service Times in a Queue. In Example 5.7.5, the conditional mean service rate given the observations $X_1 = x_1, \ldots, X_n = x_n$ is

$$E(Z|x_1,\ldots,x_n) = \frac{n+1}{2+\sum_{i=1}^n x_i}.$$

For large n, the conditional mean is approximately 1 over the sample average of the service times. This makes sense since 1 over the average service time is what we generally mean by service rate. ◂

The m.g.f. ψ of X can be obtained similarly.

Theorem 5.7.6 Moment Generating Function. Let X have the gamma distribution with parameters α and β. The m.g.f. of X is

$$\psi(t) = \left(\frac{\beta}{\beta-t}\right)^\alpha \quad \text{for } t < \beta. \tag{5.7.15}$$

Proof The m.g.f. is

$$\psi(t) = \int_0^\infty e^{tx} f(x|\alpha, \beta)\, dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\beta-t)x}\, dx.$$

This integral will be finite for every value of t such that $t < \beta$. Therefore, it follows from Eq. (5.7.10) that, for $t < \beta$,

$$\psi(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha)}{(\beta - t)^\alpha} = \left(\frac{\beta}{\beta - t}\right)^\alpha. \qquad \blacksquare$$

We can now show that the sum of independent random variables that have gamma distributions with a common value of the parameter β will also have a gamma distribution.

Theorem 5.7.7 If the random variables $X_1, \ldots, X_k$ are independent, and if X_i has the gamma distribution with parameters α_i and β $(i = 1, \ldots, k)$, then the sum $X_1 + \cdots + X_k$ has the gamma distribution with parameters $\alpha_1 + \cdots + \alpha_k$ and β.

Proof If ψ_i denotes the m.g.f. of X_i, then it follows from Eq. (5.7.15) that for $i = 1, \ldots, k$,

$$\psi_i(t) = \left(\frac{\beta}{\beta - t}\right)^{\alpha_i} \quad \text{for } t < \beta.$$

If ψ denotes the m.g.f. of the sum $X_1 + \cdots + X_k$, then by Theorem 4.4.4,

$$\psi(t) = \prod_{i=1}^{k} \psi_i(t) = \left(\frac{\beta}{\beta - t}\right)^{\alpha_1 + \cdots + \alpha_k} \quad \text{for } t < \beta.$$

The m.g.f. ψ can now be recognized as the m.g.f. of the gamma distribution with parameters $\alpha_1 + \cdots + \alpha_k$ and β. Hence, the sum $X_1 + \cdots + X_k$ must have this gamma distribution. $\blacksquare$

The Exponential Distributions

A special case of gamma distributions provide a common model for phenomena such as waiting times. For instance, in Example 5.7.3, the conditional distribution of each service time X_i given Z (the rate of service) is a member of the following family of distributions.

Definition 5.7.3 Exponential Distributions. Let $\beta > 0$. A random variable X has the *exponential distribution with parameter* β if X has a continuous distribution with the p.d.f.

$$f(x|\beta) = \begin{cases} \beta e^{-\beta x} & \text{for } x > 0, \\ 0 & \text{for } x \le 0. \end{cases} \tag{5.7.16}$$

A comparison of the p.d.f.'s for gamma and exponential distributions makes the following result obvious.

Theorem 5.7.8 The exponential distribution with parameter β is the same as the gamma distribution with parameters $\alpha = 1$ and β. If X has the exponential distribution with parameter β, then

$$E(X) = \frac{1}{\beta} \quad \text{and} \quad \text{Var}(X) = \frac{1}{\beta^2}, \tag{5.7.17}$$

and the m.g.f. of X is

$$\psi(t) = \frac{\beta}{\beta - t} \quad \text{for } t < \beta. \qquad \blacksquare$$

Exponential distributions have a memoryless property similar to that stated in Theorem 5.5.5 for geometric distributions.

Theorem 5.7.9 Memoryless Property of Exponential Distributions. Let X have the exponential distribution with parameter β, and let $t > 0$. Then for every number $h > 0$,

$$\Pr(X \geq t + h | X \geq t) = \Pr(X \geq h). \tag{5.7.18}$$

Proof For each $t > 0$,

$$\Pr(X \geq t) = \int_t^\infty \beta e^{-\beta x}\, dx = e^{-\beta t}. \tag{5.7.19}$$

Hence, for each $t > 0$ and each $h > 0$,

$$\begin{aligned}\Pr(X \geq t + h | X \geq t) &= \frac{\Pr(X \geq t + h)}{\Pr(X \geq t)} \\ &= \frac{e^{-\beta(t+h)}}{e^{-\beta t}} = e^{-\beta h} = \Pr(X \geq h).\end{aligned} \tag{5.7.20}$$

$\blacksquare$

You can prove (see Exercise 23) that the exponential distributions are the only continuous distributions with the memoryless property.

To illustrate the memoryless property, we shall suppose that X represents the number of minutes that elapse before some event occurs. According to Eq. (5.7.20), if the event has not occurred during the first t minutes, then the probability that the event will not occur during the next h minutes is simply $e^{-\beta h}$. This is the same as the probability that the event would not occur during an interval of h minutes starting from time 0. In other words, regardless of the length of time that has elapsed without the occurrence of the event, the probability that the event will occur during the next h minutes always has the same value.

This memoryless property will not strictly be satisfied in all practical problems. For example, suppose that X is the length of time for which a light bulb will burn before it fails. The length of time for which the bulb can be expected to continue to burn in the future will depend on the length of time for which it has been burning in the past. Nevertheless, the exponential distribution has been used effectively as an approximate distribution for such variables as the lengths of the lives of various products.

Life Tests

Example 5.7.7 Light Bulbs. Suppose that n light bulbs are burning simultaneously in a test to determine the lengths of their lives. We shall assume that the n bulbs burn independently of one another and that the lifetime of each bulb has the exponential distribution with parameter β. In other words, if X_i denotes the lifetime of bulb i, for $i = 1, \dots, n$, then it is assumed that the random variables $X_1, \dots, X_n$ are i.i.d. and that each has the exponential distribution with parameter β. What is the distribution of the length of time Y_1 until the first failure of one of the n bulbs? What is the distribution of the length of time Y_2 after the first failure until a second bulb fails? ◀

The random variable Y_1 in Example 5.7.7 is the minimum of a random sample of n exponential random variables. The distribution of Y_1 is easy to find.

Theorem 5.7.10 Suppose that the variables $X_1, \ldots, X_n$ form a random sample from the exponential distribution with parameter β. Then the distribution of $Y_1 = \min\{X_1, \ldots, X_n\}$ will be the exponential distribution with parameter $n\beta$.

Proof For every number $t > 0$,

$$\begin{aligned}\Pr(Y_1 > t) &= \Pr(X_1 > t, \ldots, X_n > t)\\ &= \Pr(X_1 > t) \cdots \Pr(X_n > t)\\ &= e^{-\beta t} \cdots e^{-\beta t} = e^{-n\beta t}.\end{aligned}$$

By comparing this result with Eq. (5.7.19), we see that the distribution of Y_1 must be the exponential distribution with parameter $n\beta$. ■

The memoryless property of the exponential distributions allows us to answer the second question at the end of Example 5.7.7, as well as similar questions about later failures. After one bulb has failed, $n - 1$ bulbs are still burning. Furthermore, regardless of the time at which the first bulb failed or which bulb failed first, it follows from the memoryless property of the exponential distribution that the distribution of the remaining lifetime of each of the other $n - 1$ bulbs is still the exponential distribution with parameter β. In other words, the situation is the same as it would be if we were starting the test over again from time $t = 0$ with $n - 1$ new bulbs. Therefore, Y_2 will be equal to the smallest of $n - 1$ i.i.d. random variables, each of which has the exponential distribution with parameter β. It follows from Theorem 5.7.10 that Y_2 will have the exponential distribution with parameter $(n - 1)\beta$. The next result deals with the remaining waiting times between failures.

Theorem 5.7.11 Suppose that the variables $X_1, \ldots, X_n$ form a random sample from the exponential distribution with parameter β. Let $Z_1 \le Z_2 \le \cdots \le Z_n$ be the random variables $X_1, \ldots, X_n$ sorted from smallest to largest. For each $k = 2, \ldots, n$, let $Y_k = Z_k - Z_{k-1}$. Then the distribution of Y_k is the exponential distribution with parameter $(n + 1 - k)\beta$.

Proof At the time Z_{k-1}, exactly $k - 1$ of the lifetimes have ended and there are $n + 1 - k$ lifetimes that have not yet ended. For each of the remaining lifetimes, the conditional distribution of what remains of that lifetime given that it has lasted at least Z_{k-1} is still exponential with parameter β by the memoryless property. So, $Y_k = Z_k - Z_{k-1}$ has the same distribution as the minimum lifetime from a random sample of size $n + 1 - k$ from the exponential distribution with parameter β. According to Theorem 5.7.10, that distribution is exponential with parameter $(n + 1 - k)\beta$. ■

Relation to the Poisson Process

Example 5.7.8 Radioactive Particles. Suppose that radioactive particles strike a target according to a Poisson process with rate β, as defined in Definition 5.4.2. Let Z_k be the time until the kth particle strikes the target for $k = 1, 2, \ldots$. What is the distribution of Z_1? What is the distribution of $Y_k = Z_k - Z_{k-1}$ for $k \ge 2$? ◄

Although the random variables defined at the end of Example 5.7.8 look similar to those in Theorem 5.7.11, there are major differences. In Theorem 5.7.11, we were

observing a fixed number n of lifetimes that all started simultaneously. The n lifetimes are all labeled in advance, and each could be observed independently of the others. In Example 5.7.8, there is no fixed number of particles being contemplated, and we have no well-defined notion of when each particle "starts" toward the target. In fact, we cannot even tell which particle is which until after they are observed. We merely start observing at an arbitrary time and record each time a particle hits. Depending on how long we observe the process, we could see an arbitrary number of particles hit the target in Example 5.7.8, but we could never see more than n failures in the setup of Theorem 5.7.11, no matter how long we observe. Theorem 5.7.12 gives the distributions for the times between arrivals in Example 5.7.8, and one can see how the distributions differ from those in Theorem 5.7.11.

Theorem 5.7.12 Times between Arrivals in a Poisson Process. Suppose that arrivals occur according to a Poisson process with rate β. Let Z_k be the time until the kth arrival for $k = 1, 2, \dots$. Define $Y_1 = Z_1$ and $Y_k = Z_k - Z_{k-1}$ for $k \geq 2$. Then $Y_1, Y_2, \dots$ are i.i.d. and they each have the exponential distribution with parameter β.

Proof Let $t > 0$, and define X to be the number of arrivals from time 0 until time t. It is easy to see that $Y_1 \leq t$ if and only if $X \geq 1$. That is, the first particle strikes the target by time t if and only if at least one particle strikes the target by time t. We already know that X has the Poisson distribution with mean βt, where β is the rate of the process. So, for $t > 0$,

$$\Pr(Y_1 \leq t) = \Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - e^{-\beta t}.$$

Comparing this to Eq. (5.7.19), we see that $1 - e^{-\beta t}$ is the c.d.f. of the exponential distribution with parameter β.

What happens in a Poisson process after time t is independent of what happens up to time t. Hence, the conditional distribution given $Y_1 = t$ of the gap from time t until the next arrival at Z_2 is the same as the distribution of the time from time 0 until the first arrival. That is, the distribution of $Y_2 = Z_2 - Z_1$ given $Y_1 = t$ (i.e., $Z_1 = t$) is the exponential distribution with parameter β no matter what t is. Hence, Y_2 is independent of Y_1 and they have the same distribution. The same argument can be applied to find the distributions for $Y_3, Y_4, \dots$. ■

An exponential distribution is often used in a practical problem to represent the distribution of the time that elapses before the occurrence of some event. For example, this distribution has been used to represent such periods of time as the period for which a machine or an electronic component will operate without breaking down, the period required to take care of a customer at some service facility, and the period between the arrivals of two successive customers at a facility.

If the events being considered occur in accordance with a Poisson process, then both the waiting time until an event occurs and the period of time between any two successive events will have exponential distributions. This fact provides theoretical support for the use of the exponential distribution in many types of problems.

We can combine Theorem 5.7.12 with Theorem 5.7.7 to obtain the following.

Corollary 5.7.1 Time until kth Arrival. In the situation of Theorem 5.7.12, the distribution of Z_k is the gamma distribution with parameters k and β. ■

Summary

The gamma function is defined by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$ and has the property that $\Gamma(n) = (n-1)!$ for $n = 1, 2, \ldots$. If $X_1, \ldots, X_n$ are independent random variables with gamma distributions all having the same second parameter β, then $\sum_{i=1}^n X_i$ has the gamma distribution with first parameter equal to the sum of the first parameters of $X_1, \ldots, X_n$ and second parameter equal to β. The exponential distribution with parameter β is the same as the gamma distribution with parameters 1 and β. Hence, the sum of a random sample of n exponential random variables with parameter β has the gamma distribution with parameters n and β. For a Poisson process with rate β, the times between successive occurrences have the exponential distribution with parameter β, and they are independent. The waiting time until the kth occurrence has the gamma distribution with parameters k and β.

Exercises

1. Suppose that X has the gamma distribution with parameters α and β, and c is a positive constant. Show that cX has the gamma distribution with parameters α and β/c.

2. Compute the quantile function of the exponential distribution with parameter β.

3. Sketch the p.d.f. of the gamma distribution for each of the following pairs of values of the parameters α and β: **(a)** $\alpha = 1/2$ and $\beta = 1$, **(b)** $\alpha = 1$ and $\beta = 1$, **(c)** $\alpha = 2$ and $\beta = 1$.

4. Determine the mode of the gamma distribution with parameters α and β.

5. Sketch the p.d.f. of the exponential distribution for each of the following values of the parameter β: **(a)** $\beta = 1/2$, **(b)** $\beta = 1$, and **(c)** $\beta = 2$.

6. Suppose that $X_1, \ldots, X_n$ form a random sample of size n from the exponential distribution with parameter β. Determine the distribution of the sample mean $\overline{X}_n$.

7. Let X_1, X_2, X_3 be a random sample from the exponential distribution with parameter β. Find the probability that at least one of the random variables is greater than t, where $t > 0$.

8. Suppose that the random variables $X_1, \ldots, X_k$ are independent and X_i has the exponential distribution with parameter β_i $(i = 1, \ldots, k)$. Let $Y = \min\{X_1, \ldots, X_k\}$. Show that Y has the exponential distribution with parameter $\beta_1 + \cdots + \beta_k$.

9. Suppose that a certain system contains three components that function independently of each other and are connected in series, as defined in Exercise 5 of Sec. 3.7, so that the system fails as soon as one of the components fails. Suppose that the length of life of the first component, measured in hours, has the exponential distribution with parameter $\beta = 0.001$, the length of life of the second component has the exponential distribution with parameter $\beta = 0.003$, and the length of life of the third component has the exponential distribution with parameter $\beta = 0.006$. Determine the probability that the system will not fail before 100 hours.

10. Suppose that an electronic system contains n similar components that function independently of each other and that are connected in series so that the system fails as soon as one of the components fails. Suppose also that the length of life of each component, measured in hours, has the exponential distribution with mean μ. Determine the mean and the variance of the length of time until the system fails.

11. Suppose that n items are being tested simultaneously, the items are independent, and the length of life of each item has the exponential distribution with parameter β. Determine the expected length of time until three items have failed. *Hint:* The required value is $E(Y_1 + Y_2 + Y_3)$ in the notation of Theorem 5.7.11.

12. Consider again the electronic system described in Exercise 10, but suppose now that the system will continue to operate until two components have failed. Determine the mean and the variance of the length of time until the system fails.

13. Suppose that a certain examination is to be taken by five students independently of one another, and the number of minutes required by any particular student to complete the examination has the exponential distribution for which the mean is 80. Suppose that the examination begins at 9:00 A.M. Determine the probability that at least one of the students will complete the examination before 9:40 A.M.

14. Suppose again that the examination considered in Exercise 13 is taken by five students, and the first student to complete the examination finishes at 9:25 A.M. Determine the probability that at least one other student will complete the examination before 10:00 A.M.

15. Suppose again that the examination considered in Exercise 13 is taken by five students. Determine the probability that no two students will complete the examination within 10 minutes of each other.

16. It is said that a random variable X has the *Pareto distribution with parameters* x_0 *and* α ($x_0 > 0$ and $\alpha > 0$) if X has a continuous distribution for which the p.d.f. $f(x|x_0, \alpha)$ is as follows:

$$f(x|x_0, \alpha) = \begin{cases} \dfrac{\alpha x_0^\alpha}{x^{\alpha+1}} & \text{for } x \geq x_0, \\ 0 & \text{for } x < x_0. \end{cases}$$

Show that if X has this Pareto distribution, then the random variable $\log(X/x_0)$ has the exponential distribution with parameter α.

17. Suppose that a random variable X has the normal distribution with mean μ and variance σ^2. Determine the value of $E[(X - \mu)^{2n}]$ for $n = 1, 2, \ldots$.

18. Consider a random variable X for which $\Pr(X > 0) = 1$, the p.d.f. is f, and the c.d.f. is F. Consider also the function h defined as follows:

$$h(x) = \frac{f(x)}{1 - F(x)} \quad \text{for } x > 0.$$

The function h is called the *failure rate* or the *hazard function* of X. Show that if X has an exponential distribution, then the failure rate $h(x)$ is constant for $x > 0$.

19. It is said that a random variable has the *Weibull distribution with parameters* a *and* b ($a > 0$ and $b > 0$) if X has a continuous distribution for which the p.d.f. $f(x|a, b)$ is as follows:

$$f(x|a, b) = \begin{cases} \dfrac{b}{a^b} x^{b-1} e^{-(x/a)^b} & \text{for } x > 0, \\ 0 & \text{for } x \leq 0. \end{cases}$$

Show that if X has this Weibull distribution, then the random variable X^b has the exponential distribution with parameter $\beta = a^{-b}$.

20. It is said that a random variable X has an *increasing failure rate* if the failure rate $h(x)$ defined in Exercise 18 is an increasing function of x for $x > 0$, and it is said that X has a *decreasing failure rate* if $h(x)$ is a decreasing function of x for $x > 0$. Suppose that X has the Weibull distribution with parameters a and b, as defined in Exercise 19. Show that X has an increasing failure rate if $b > 1$, and X has a decreasing failure rate if $b < 1$.

21. Let X have the gamma distribution with parameters $\alpha > 2$ and $\beta > 0$.

- **a.** Prove that the mean of $1/X$ is $\beta/(\alpha - 1)$.
- **b.** Prove that the variance of $1/X$ is $\beta^2/[(\alpha - 1)^2 (\alpha - 2)]$.

22. Consider the Poisson process of radioactive particle hits in Example 5.7.8. Suppose that the rate β of the Poisson process is unknown and has the gamma distribution with parameters α and γ. Let X be the number of particles that strike the target during t time units. Prove that the conditional distribution of β given $X = x$ is a gamma distribution, and find the parameters of that gamma distribution.

23. Let F be a continuous c.d.f. satisfying $F(0) = 0$, and suppose that the distribution with c.d.f. F has the memoryless property (5.7.18). Define $\ell(x) = \log[1 - F(x)]$ for $x > 0$.

- **a.** Show that for all $t, h > 0$,
$$1 - F(h) = \frac{1 - F(t + h)}{1 - F(t)}.$$
- **b.** Prove that $\ell(t + h) = \ell(t) + \ell(h)$ for all $t, h > 0$.
- **c.** Prove that for all $t > 0$ and all positive integers k and m, $\ell(kt/m) = (k/m)\ell(t)$.
- **d.** Prove that for all $t, c > 0$, $\ell(ct) = c\ell(t)$.
- **e.** Prove that $g(t) = \ell(t)/t$ is constant for $t > 0$.
- **f.** Prove that F must be the c.d.f. of an exponential distribution.

24. Review the derivation of the Black-Scholes formula (5.6.18). For this exercise, assume that our stock price at time u in the future is $S_0 e^{\mu u + W_u}$, where W_u has the gamma distribution with parameters αu and β with $\beta > 1$. Let r be the risk-free interest rate.

- **a.** Prove that $e^{-ru}E(S_u) = S_0$ if and only if $\mu = r - \alpha \log(\beta/[\beta - 1])$.
- **b.** Assume that $\mu = r - \alpha \log(\beta/[\beta - 1])$. Let R be 1 minus the c.d.f. of the gamma distribution with parameters αu and 1. Prove that the risk-neutral price for the option to buy one share of the stock for the price q at time u is $S_0 R(c[\beta - 1]) - q e^{-ru} R(c\beta)$, where
$$c = \log\left(\frac{q}{S_0}\right) + \alpha u \log\left(\frac{\beta}{\beta - 1}\right) - ru.$$
- **c.** Find the price for the option being considered when $u = 1$, $q = S_0$, $r = 0.06$, $\alpha = 1$, and $\beta = 10$.

5.8 The Beta Distributions

The family of beta distributions is a popular model for random variables that are known to take values in the interval [0, 1]. *One common example of such a random variable is the unknown proportion of successes in a sequence of Bernoulli trials.*

The Beta Function

Example 5.8.1

Defective Parts. A machine produces parts that are either defective or not, as in Example 3.6.9 on page 148. Let P denote the proportion of defectives among all parts that might be produced by this machine. Suppose that we observe n such parts, and let X be the number of defectives among the n parts observed. If we assume that the parts are conditionally independent given P, then we have the same situation as in Example 3.6.9, where we computed the conditional p.d.f. of P given $X = x$ as

$$g_2(p|x) = \frac{p^x(1-p)^{n-x}}{\int_0^1 q^x(1-q)^{n-x}dq}, \quad \text{for } 0 < p < 1. \tag{5.8.1}$$

We are now in a position to calculate the integral in the denominator of Eq. (5.8.1). The distribution with the resulting p.d.f. is a member a useful family that we shall study in this section. ◀

Definition 5.8.1

The Beta Function. For each positive α and β, define

$$\mathrm{B}(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

The function B is called the *beta function*.

We can show that the beta function B is finite for all $\alpha, \beta > 0$. The proof of the following result relies on the methods from the end of Sec. 3.9 and is given at the end of this section.

Theorem 5.8.1

For all $\alpha, \beta > 0$,

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{5.8.2}$$

Example 5.8.2

Defective Parts. It follows from Theorem 5.8.1 that the integral in the denominator of Eq. (5.8.1) is

$$\int_0^1 q^x(1-q)^{n-x}dq = \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} = \frac{x!(n-x)!}{(n+1)!}.$$

The conditional p.d.f. of P given $X = x$ is then

$$g_2(p|x) = \frac{(n+1)!}{x!(n-x)!}p^x(1-p)^{n-x}, \text{ for } 0 < p < 1.$$ ◀

Definition of the Beta Distributions

The distribution in Example 5.8.2 is a special case of the following.

Definition 5.8.2 Beta Distributions. Let $\alpha, \beta > 0$ and let X be a random variable with p.d.f.

$$f(x|\alpha,\beta) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{5.8.3}$$

Then X has the *beta distribution with parameters α and β.*

The conditional distribution of P given $X = x$ in Example 5.8.2 is the beta distribution with parameters $x + 1$ and $n - x + 1$. It can also be seen from Eq. (5.8.3) that the beta distribution with parameters $\alpha = 1$ and $\beta = 1$ is simply the uniform distribution on the interval [0, 1].

Example 5.8.3 Castaneda v. Partida. In Example 5.2.6 on page 278, 220 grand jurors were chosen from a population that is 79.1 percent Mexican American, but only 100 grand jurors were Mexican American. The expected value of a binomial random variable X with parameters 220 and 0.791 is $E(X) = 220 \times 0.791 = 174.02$. This is much larger than the observed value of $X = 100$. Of course, such a discrepancy could occur by chance. After all, there is positive probability of $X = x$ for all $x = 0, \dots, 220$. Let P stand for the proportion of Mexican Americans among all grand jurors that would be chosen under the current system being used. The court assumed that X had the binomial distribution with parameters $n = 220$ and p, conditional on $P = p$. We should then be interested in whether P is substantially less than the value 0.791, which represents impartial juror choice. For example, suppose that we define discrimination to mean that $P \le 0.8 \times 0.791 = 0.6328$. We would like to compute the conditional probability of $P \le 0.6328$ given $X = 100$.

Suppose that the distribution of P prior to observing X was the beta distribution with parameters α and β. Then the p.d.f. of P was

$$f_2(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad \text{for } 0 < p < 1.$$

The conditional p.f. of X given $P = p$ is the binomial p.f.

$$g_1(x|p) = \binom{220}{x} p^x (1-p)^{220-x}, \quad \text{for } x = 0, \dots, 220.$$

We can now apply Bayes' theorem for random variables (3.6.13) to obtain the conditional p.d.f. of P given $X = 100$:

$$\begin{aligned} g_2(p|100) &= \frac{\binom{220}{100} p^{100}(1-p)^{120} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{f_1(100)} \\ &= \frac{\binom{220}{100}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta) f_1(100)} p^{\alpha+100-1}(1-p)^{\beta+120-1}, \end{aligned} \tag{5.8.4}$$

for $0 < p < 1$, where $f_1(100)$ is the marginal p.f. of X at 100. As a function of p the far right side of Eq. (5.8.4) is a constant times $p^{\alpha+100-1}(1-p)^{\beta+120-1}$ for $0 < p < 1$. As such, it is clearly the p.d.f. of a beta distribution. The parameters

of that beta distribution are $\alpha + 100$ and $\beta + 120$. Hence, the constant must be $1/\mathrm{B}(100 + \alpha, 120 + \beta)$. That is,

$$g_2(p|100) = \frac{\Gamma(\alpha + \beta + 220)}{\Gamma(\alpha + 100)\Gamma(\beta + 120)} p^{\alpha+100-1}(1 - p)^{\beta+120-1}, \quad \text{for } 0 < p < 1. \tag{5.8.5}$$

After choosing values of α and β, we could compute $\Pr(P \le 0.6328|X = 100)$ and decide how likely it is that there was discrimination. We will see how to choose α and β after we learn how to compute the expected value of a beta random variable. ◀

Note: Conditional Distribution of P after Observing X with Binomial Distribution. The calculation of the conditional distribution of P given $X = 100$ in Example 5.8.3 is a special case of a useful general result. In fact, the proof of the following result is essentially given in Example 5.8.3, and will not be repeated.

Theorem 5.8.2 Suppose that P has the beta distribution with parameters α and β, and the conditional distribution of X given $P = p$ is the binomial distribution with parameters n and p. Then the conditional distribution of P given $X = x$ is the beta distribution with parameters $\alpha + x$ and $\beta + n - x$. ■

Moments of Beta Distributions

Theorem 5.8.3 Moments. Suppose that X has the beta distribution with parameters α and β. Then for each positive integer k,

$$E(X^k) = \frac{\alpha(\alpha + 1) \cdots (\alpha + k - 1)}{(\alpha + \beta)(\alpha + \beta + 1) \cdots (\alpha + \beta + k - 1)}. \tag{5.8.6}$$

In particular,

$$E(X) = \frac{\alpha}{\alpha + \beta},$$

$$\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Proof For $k = 1, 2, \ldots,$

$$E(X^k) = \int_0^1 x^k f(x|\alpha, \beta)\, dx$$

$$= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+k-1}(1 - x)^{\beta-1}\, dx.$$

Therefore, by Eq. (5.8.2),

$$E(X^k) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha + k)\Gamma(\beta)}{\Gamma(\alpha + k + \beta)},$$

which simplifies to Eq. (5.8.6). The special case of the mean is simple, while the variance follows easily from

$$E(X^2) = \frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)}.$$ ■

There are too many beta distributions to provide tables in the back of the book. Any good statistical package will be able to calculate the c.d.f.'s of many beta

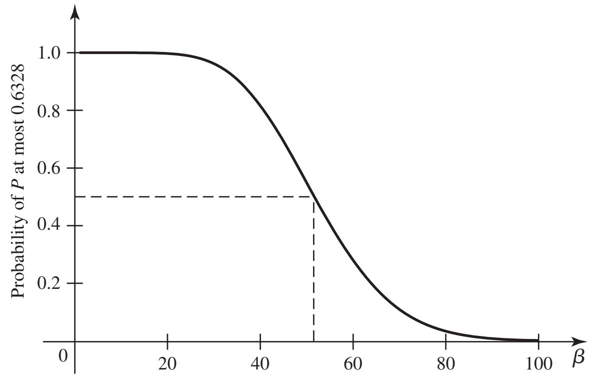


Figure 5.8 Probability of discrimination as a function of β.

distributions, and some packages will also be able to calculate the quantile functions. The next example illustrates the importance of being able to calculate means and c.d.f.'s of beta distributions.

Example 5.8.4

Castaneda v. Partida. Continuing Example 5.8.3, we are now prepared to see why, for every reasonable choice one makes for α and β, the probability of discrimination in Castaneda v. Partida is quite large. To avoid bias either for or against the defendant, we shall suppose that, before learning X, the probability that a Mexican American juror would be selected on each draw from the pool was 0.791. Let $Y = 1$ if a Mexican American juror is selected on a single draw, and let $Y = 0$ if not. Then Y has the Bernoulli distribution with parameter p given $P = p$ and $E(Y|p) = p$. So the law of total probability for expectations, Theorem 4.7.1, says that

$$\Pr(Y = 1) = E(Y) = E[E(Y|P)] = E(P).$$

This means that we should choose α and β so that $E(P) = 0.791$. Because $E(P) = \alpha/(\alpha + \beta)$, this means that $\alpha = 3.785\beta$. The conditional distribution of P given $X = 100$ is the beta distribution with parameters $\alpha + 100$ and $\beta + 120$. For each value of $\beta > 0$, we can compute $\Pr(P \le 0.6328|X = 100)$ using $\alpha = 3.785\beta$. Then, for each β we can check whether or not that probability is small. A plot of $\Pr(P \le 0.6328|X = 100)$ for various values of β is given in Fig. 5.8. From the figure, we see that $\Pr(P \le 0.6328|X = 100) < 0.5$ only for $\beta \ge 51.5$. This makes $\alpha \ge 194.9$. We claim that the beta distribution with parameters 194.9 and 51.5 as well as all others that make $\Pr(P \le 0.6328|X = 100) < 0.5$ are unreasonable because they are incredibly prejudiced about the possibility of discrimination. For example, suppose that someone actually believed, before observing $X = 100$, that the distribution of P was the beta distribution with parameters 194.9 and 51.5. For this beta distribution, the probability that there is discrimination would be $\Pr(P \le 0.6328) = 3.28 \times 10^{-8}$, which is essentially 0. All of the other priors with $\beta \ge 51.5$ and $\alpha = 3.785\beta$ have even smaller probabilities of $\{P \le 0.6328\}$. Arguing from the other direction, we have the following: Anyone who believed, before observing $X = 100$, that $E(P) = 0.791$ and the probability of discrimination was greater than 3.28×10^{-8}, would believe that the probability of discrimination is at least 0.5 after learning $X = 100$. This is then fairly convincing evidence that there was discrimination in this case. ◀

Example 5.8.5

A Clinical Trial. Consider the clinical trial described in Example 2.1.4. Let P be the proportion of all patients in a large group receiving imipramine who have no relapse (called success). A popular model for P is that P has the beta distribution with

parameters α and β. Choosing α and β can be done based on expert opinion about the chance of success and on the effect that data should have on the distribution of P after observing the data. For example, suppose that the doctors running the clinical trial think that the probability of success should be around 1/3. Let $X_i = 1$ if the ith patient is a success and $X_i = 0$ if not. We are supposing that $E(X_i|p) = \Pr(X_i = 1|p) = p$, so the law of total probability for expectations (Theorem 4.7.1) says that

$$\Pr(X_i = 1) = E(X_i) = E[E(X_i|P)] = E(P) = \frac{\alpha}{\alpha + \beta}.$$

If we want $\Pr(X_i = 1) = 1/3$, we need $\alpha/(\alpha + \beta) = 1/3$, so $\beta = 2\alpha$. Of course, the doctors will revise the probability of success after observing patients from the study. The doctors can choose α and β based on how that revision will occur.

Assume that the random variables $X_1, X_2, \ldots$ (the indicators of success) are conditionally independent given $P = p$. Let $X = X_1 + \cdots + X_n$ be the number of patients out of the first n who are successes. The conditional distribution of X given $P = p$ is the binomial distribution with parameters n and p, and the marginal distribution of P is the beta distribution with parameters α and β. Theorem 5.8.2 tells us that the conditional distribution of P given $X = x$ is the beta distribution with parameters $\alpha + x$ and $\beta + n - x$. Suppose that a sequence of 20 patients, all of whom are successes, would raise the doctors' probability of success from 1/3 up to 0.9. Then

$$0.9 = E(P|X = 20) = \frac{\alpha + 20}{\alpha + \beta + 20}.$$

This equation implies that $\alpha + 20 = 9\beta$. Combining this with $\beta = 2\alpha$, we get $\alpha = 1.18$ and $\beta = 2.35$.

Finally, we can ask, what will be the distribution of P after observing some patients in the study? Suppose that 40 patients are actually observed, and 22 of them recover (as in Table 2.1). Then the conditional distribution of P given this observation is the beta distribution with parameters $1.18 + 22 = 23.18$ and $2.35 + 18 = 20.35$. It follows that

$$E(P|X = 22) = \frac{23.18}{23.18 + 20.35} = 0.5325.$$

Notice how much closer this is to the proportion of successes (0.55) than was $E(P) = 1/3$. ◄

Proof of Theorem 5.8.1.

Theorem 5.8.1, i.e., Eq. (5.8.2), is part of the following useful result. The proof uses Theorem 3.9.5 (multivariate transformation of random variables). If you did not study Theorem 3.9.5, you will not be able to follow the proof of Theorem 5.8.4.

Theorem 5.8.4 Let U and V be independent random variables with U having the gamma distribution with parameters α and 1 and V having the gamma distribution with parameters β and 1. Then

- $X = U/(U + V)$ and $Y = U + V$ are independent,
- X has the beta distribution with parameters α and β, and
- Y has the gamma distribution with parameters $\alpha + \beta$ and 1.

Also, Eq. (5.8.2) holds.

Proof Because U and V are independent, the joint p.d.f. of U and V is the product of their marginal p.d.f.'s, which are

$$f_1(u) = \frac{u^{\alpha-1}e^{-u}}{\Gamma(\alpha)}, \text{ for } u > 0,$$

$$f_2(v) = \frac{v^{\beta-1}e^{-v}}{\Gamma(\beta)}, \text{ for } v > 0.$$

So, the joint p.d.f. is

$$f(u, v) = \frac{u^{\alpha-1}v^{\beta-1}e^{-(u+v)}}{\Gamma(\alpha)\Gamma(\beta)},$$

for $u > 0$ and $v > 0$.

The transformation from (u, v) to (x, y) is

$$x = r_1(u, v) = \frac{u}{u+v} \quad \text{and} \quad y = r_2(u, v) = u + v,$$

and the inverse is

$$u = s_1(x, y) = xy \text{ and } v = s_2(x, y) = (1 - x)y.$$

The Jacobian is the determinant of the matrix

$$J = \begin{bmatrix} y & x \\ -y & 1-x \end{bmatrix},$$

which equals y. According to Theorem 3.9.5, the joint p.d.f. of (X, Y) is then

$$\begin{aligned} g(x, y) &= f(s_1(x, y), s_2(x, y))y \\ &= \frac{x^{\alpha-1}(1-x)^{\beta-1}y^{\alpha+\beta-1}e^{-y}}{\Gamma(\alpha)\Gamma(\beta)}, \end{aligned} \tag{5.8.7}$$

for $0 < x < 1$ and $y > 0$. Notice that this joint p.d.f. factors into separate functions of x and y, and hence X and Y are independent. The marginal distribution of Y is available from Theorem 5.7.7. The marginal p.d.f. of X is obtained by integrating y out of (5.8.7):

$$\begin{aligned} g_1(x) &= \int_0^\infty \frac{x^{\alpha-1}(1-x)^{\beta-1}y^{\alpha+\beta-1}e^{-y}}{\Gamma(\alpha)\Gamma(\beta)}dy \\ &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty y^{\alpha+\beta-1}e^{-y}dy \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \end{aligned} \tag{5.8.8}$$

where the last equation follows from (5.7.2). Because the far right side of (5.8.8) is a p.d.f., it integrates to 1, which proves Eq. (5.8.2). Also, one can recognize the far right side of (5.8.8) as the p.d.f. of the beta distribution with parameters α and β. ■

Summary

The family of beta distributions is a popular model for random variables that lie in the interval (0, 1), such as unknown proportions of success for sequences of Bernoulli trials. The mean of the beta distribution with parameters α and β is $\alpha/(\alpha + \beta)$. If X

has the binomial distribution with parameters n and p conditional on $P = p$, and if P has the beta distribution with parameters α and β, then, conditional on $X = x$, the distribution of P is the beta distribution with parameters $\alpha + x$ and $\beta + n - x$.

Exercises

1. Compute the quantile function of the beta distribution with parameters $\alpha > 0$ and $\beta = 1$.

2. Determine the mode of the beta distribution with parameters α and β, assuming that $\alpha > 1$ and $\beta > 1$.

3. Sketch the p.d.f. of the beta distribution for each of the following pairs of values of the parameters:

a. $\alpha = 1/2$ and $\beta = 1/2$ **b.** $\alpha = 1/2$ and $\beta = 1$
c. $\alpha = 1/2$ and $\beta = 2$ **d.** $\alpha = 1$ and $\beta = 1$
e. $\alpha = 1$ and $\beta = 2$ **f.** $\alpha = 2$ and $\beta = 2$
g. $\alpha = 25$ and $\beta = 100$ **h.** $\alpha = 100$ and $\beta = 25$

4. Suppose that X has the beta distribution with parameters α and β. Show that $1 - X$ has the beta distribution with parameters β and α.

5. Suppose that X has the beta distribution with parameters α and β, and let r and s be given positive integers. Determine the value of $E[X^r(1 - X)^s]$.

6. Suppose that X and Y are independent random variables, X has the gamma distribution with parameters α_1 and β, and Y has the gamma distribution with parameters α_2 and β. Let $U = X/(X + Y)$ and $V = X + Y$. Show that **(a)** U has the beta distribution with parameters α_1 and α_2, and **(b)** U and V are independent. *Hint:* Look at the steps in the proof of Theorem 5.8.1.

7. Suppose that X_1 and X_2 form a random sample of two observed values from the exponential distribution with parameter β. Show that $X_1/(X_1 + X_2)$ has the uniform distribution on the interval [0, 1].

8. Suppose that the proportion X of defective items in a large lot is unknown and that X has the beta distribution with parameters α and β.

a. If one item is selected at random from the lot, what is the probability that it will be defective?

b. If two items are selected at random from the lot, what is the probability that both will be defective?

9. A manufacturer believes that an unknown proportion P of parts produced will be defective. She models P as having a beta distribution. The manufacturer thinks that P should be around 0.05, but if the first 10 observed products were all defective, the mean of P would rise from 0.05 to 0.9. Find the beta distribution that has these properties.

10. A marketer is interested in how many customers are likely to buy a particular product in a particular store. Let P be the proportion of all customers in the store who will buy the product. Let the distribution of P be uniform on the interval [0, 1] before observing any data. The marketer then observes 25 customers and only six buy the product. If the customers were conditionally independent given P, find the conditional distribution of P given the observed customers.

5.9 The Multinomial Distributions

Many times we observe data that can assume three or more possible values. The family of multinomial distributions is an extension of the family of binomial distributions to handle these cases. The multinomial distributions are multivariate distributions.

Definition and Derivation of Multinomial Distributions

Example 5.9.1 Blood Types. In Example 1.8.4 on page 34, we discussed human blood types, of which there are four: O, A, B, and AB. If a number of people are chosen at random, we might be interested in the probability of obtaining certain numbers of each blood type. Such calculations are used in the courts during paternity suits. ◄

In general, suppose that a population contains items of k different types ($k \geq 2$) and that the proportion of the items in the population that are of type i is p_i

$(i = 1, \ldots, k)$. It is assumed that $p_i > 0$ for $i = 1, \ldots, k$, and $\sum_{i=1}^{k} p_i = 1$. Let $\boldsymbol{p} = (p_1, \ldots, p_k)$ denote the vector of these probabilities.

Next, suppose that n items are selected at random from the population, with replacement, and let X_i denote the number of selected items that are of type i $(i = 1, \ldots, k)$. Because the n items are selected from the population at random with replacement, the selections will be independent of each other. Hence, the probability that the first item will be of type i_1, the second item of type i_2, and so on, is simply $p_{i_1} p_{i_2} \ldots p_{i_n}$. Therefore, the probability that the sequence of n outcomes will consist of exactly x_1 items of type 1, x_2 items of type 2, and so on, selected in a *particular prespecified order*, is $p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$. It follows that the probability of obtaining exactly x_i items of type i $(i = 1, \ldots, k)$ is equal to the probability $p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$ multiplied by the total number of different ways in which the order of the n items can be specified.

From the discussion that led to the definition of multinomial coefficients (Definition 1.9.1), it follows that the total number of different ways in which n items can be arranged when there are x_i items of type i $(i = 1, \ldots, k)$ is given by the multinomial coefficient

$$\binom{n}{x_1, \ldots, x_k} = \frac{n!}{x_1! x_2! \cdots x_k!}.$$

In the notation of multivariate distributions, let $\boldsymbol{X} = (X_1, \ldots, X_k)$ denote the random vector of counts, and let $\boldsymbol{x} = (x_1, \ldots, x_k)$ denote a possible value for that random vector. Finally, let $f(\boldsymbol{x}|n, \boldsymbol{p})$ denote the joint p.f. of $\boldsymbol{X}$. Then

$$\begin{aligned} f(\boldsymbol{x}|n, \boldsymbol{p}) &= \Pr(\boldsymbol{X} = \boldsymbol{x}) = \Pr(X_1 = x_1, \ldots, X_k = x_k) \\ &= \begin{cases} \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k} & \text{if } x_1 + \cdots + x_k = n, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{5.9.1}$$

Definition 5.9.1 Multinomial Distributions. A discrete random vector $\boldsymbol{X} = (X_1, \ldots, X_k)$ whose p.f. is given by Eq. (5.9.1) has the *multinomial distribution with parameters n and* $\boldsymbol{p} = (p_1, \ldots, p_k)$.

Example 5.9.2 Attendance at a Baseball Game. Suppose that 23 percent of the people attending a certain baseball game live within 10 miles of the stadium, 59 percent live between 10 and 50 miles from the stadium, and 18 percent live more than 50 miles from the stadium. Suppose also that 20 people are selected at random from the crowd attending the game. We shall determine the probability that seven of the people selected live within 10 miles of the stadium, eight of them live between 10 and 50 miles from the stadium, and five of them live more than 50 miles from the stadium.

We shall assume that the crowd attending the game is so large that it is irrelevant whether the 20 people are selected with or without replacement. We can therefore assume that they were selected with replacement. It then follows from Eq. (5.9.1) that the required probability is

$$\frac{20!}{7!\,8!\,5!}(0.23)^7(0.59)^8(0.18)^5 = 0.0094. \quad ◂$$

Example 5.9.3 Blood Types. Berry and Geisser (1986) estimate the probabilities of the four blood types in Table 5.3 based on a sample of 6004 white Californians that was analyzed by Grunbaum et al. (1978). Suppose that we will select two people at random from this population and observe their blood types. What is the probability that they will both have the same blood type? The event that the two people have the same blood type is the union of four disjoint events, each of which is the event that the two people

Table 5.3 Estimated probabilities of blood types for white Californians

A	B	AB	O
0.360	0.123	0.038	0.479

both have one of the four different blood types. Each of these events has probability $\binom{2}{2,0,0,0}$ times the square of one of the four probabilities. The probability that we want is the sum of the probabilities of the four events:

$$\binom{2}{2,\,0,\,0,\,0}(0.360^2 + 0.123^2 + 0.038^2 + 0.479^2) = 0.376. \qquad ◀$$

Relation between the Multinomial and Binomial Distributions

When the population being sampled contains only two different types of items, that is, when $k = 2$, each multinomial distribution reduces to essentially a binomial distribution. The precise form of this relationship is as follows.

Theorem 5.9.1 Suppose that the random vector $\boldsymbol{X} = (X_1, X_2)$ has the multinomial distribution with parameters n and $\boldsymbol{p} = (p_1, p_2)$. Then X_1 has the binomial distribution with parameters n and p_1, and $X_2 = n - X_1$.

Proof It is clear from the definition of multinomial distributions that $\boldsymbol{X}_2 = n - X_1$ and $p_2 = 1 - p_1$. Therefore, the random vector $\boldsymbol{X}$ is actually determined by the single random variable X_1. From the derivation of the multinomial distribution, we see that X_1 is the number of items of type 1 that are selected if n items are selected from a population consisting of two types of items. If we call items of type 1 "success," then X_1 is the number of successes in n Bernoulli trials with probability of success on each trial equal to p_1. It follows that X_1 has the binomial distribution with parameters n and p_1. ■

The proof of Theorem 5.9.1 extends easily to the following result.

Corollary 5.9.1 Suppose that the random vector $\boldsymbol{X} = (X_1, \ldots, X_k)$ has the multinomial distribution with parameters n and $\boldsymbol{p} = (p_1, \ldots, p_k)$. The marginal distribution of each variable X_i $(i = 1, \ldots, k)$ is the binomial distribution with parameters n and p_i.

Proof Choose one i from $1, \ldots, k$, and define success to be the selection of an item of type i. Then X_i is the number of successes in n Bernoulli trials with probability of sucess on each trial equal to p_i. ■

A further generalization of Corollary 5.9.1 is that the marginal distribution of the sum of some of the coordinates of a multinomial vector has a binomial distribution. The proof is left to Exercise 1 in this section.

Corollary 5.9.2 Suppose that the random vector $\boldsymbol{X} = (X_1, \ldots, X_k)$ has the multinomial distribution with parameters n and $\boldsymbol{p} = (p_1, \ldots, p_k)$ with $k > 2$. Let $\ell < k$, and let $i_1, \ldots, i_\ell$ be distinct elements of the set $\{1, \ldots, k\}$. The distribution of $Y = X_{i_1} + \cdots + X_{i_\ell}$ is the binomial distribution with parameters n and $p_{i_1} + \cdots + p_{i_\ell}$. ■

As a final note, the relationship between Bernoulli and binomial distributions extends to multinomial distributions. The Bernoulli distribution with parameter p is the same as the binomial distribution with parameters 1 and p. However, there is no separate name for a multinomial distribution with first parameter $n = 1$. A random vector with such a distribution will consist of a single 1 in one of its coordinates and $k - 1$ zeros in the other coordinates. The probability is p_i that the ith coordinate is the 1. A k-dimensional vector seems an unwieldy way to represent a random object that can take only k different values. A more common representation would be as a single discrete random variable X that takes one of the k values $1, \dots, k$ with probabilities $p_1, \dots, p_k$, respectively. The univarite distribution just described has no famous name associated with it; however, we have just shown that it is closely related to the multinomial distribution with parameters 1 and $(p_1, \dots, p_k)$.

Means, Variances, and Covariances

The means, variances, and covaraiances of the coordinates of a multinomial random vector are given by the next result.

Theorem 5.9.2 Means, Variances, and Covariances. Let the random vector $\boldsymbol{X}$ have the multinomial distribution with parameters n and $\boldsymbol{p}$. The means and variances of the coordinates of X are

$$E(X_i) = np_i \quad \text{and} \quad \text{Var}(X_i) = np_i(1 - p_i) \quad \text{for } i = 1, \dots, k. \tag{5.9.2}$$

Also, the covariances between the coordinates are

$$\text{Cov}(X_i, X_j) = -np_i p_j. \tag{5.9.3}$$

Proof Corollary 5.9.1 says that the marginal distribution of each component X_i is the binomial distribution with parameters n and p_i. Eq. 5.9.2 follows directly from this fact.

Corollary 5.9.2 says that $X_i + X_j$ has the binomial distribution with parameters n and $p_i + p_j$. Hence,

$$\text{Var}(X_i + X_j) = n(p_i + p_j)(1 - p_i - p_j). \tag{5.9.4}$$

According to Theorem 4.6.6, it is also true that

$$\begin{aligned}\text{Var}(X_i + X_j) &= \text{Var}(X_i) + \text{Var}(X_j) + 2\,\text{Cov}(X_i, X_j)\\ &= np_i(1 - p_i) + np_j(1 - p_j) + 2\,\text{Cov}(X_i, X_j).\end{aligned} \tag{5.9.5}$$

Equate the right sides of (5.9.4) and (5.9.5), and solve for $\text{Cov}(X_i, X_j)$. The result is (5.9.3). ■

Note: Negative Covariance Is Natural for Multinomial Distributions. The negative covariance between different coordinates of a multinomial vector is natural since there are only n selections to be distributed among the k coordinates of the vector. If one of the coordinates is large, at least some of the others have to be small because the sum of the coordinates is fixed at n.

Summary

Multinomial distributions extend binomial distributions to counts of more than two possible outcomes. The ith coordinate of a vector having the multinomial distribution

with parameters n and $\boldsymbol{p} = (p_1, \ldots, p_k)$ has the binomial distribution with parameters n and p_i for $i = 1, \ldots, k$. Hence, the means and variances of the coordinates of a multinomial vector are the same as those of a binomial random variable. The covariance between the ith and jth coordinates is $-np_i p_j$.

Exercises

1. Prove Corollary 5.9.2.

2. Suppose that F is a continuous c.d.f. on the real line, and let α_1 and α_2 be numbers such that $F(\alpha_1) = 0.3$ and $F(\alpha_2) = 0.8$. If 25 observations are selected at random from the distribution for which the c.d.f. is F, what is the probability that six of the observed values will be less than α_1, 10 of the observed values will be between α_1 and α_2, and nine of the observed values will be greater than α_2?

3. If five balanced dice are rolled, what is the probability that the number 1 and the number 4 will appear the same number of times?

4. Suppose that a die is loaded so that each of the numbers 1, 2, 3, 4, 5, and 6 has a different probability of appearing when the die is rolled. For $i = 1, \ldots, 6$, let p_i denote the probability that the number i will be obtained, and suppose that $p_1 = 0.11$, $p_2 = 0.30$, $p_3 = 0.22$, $p_4 = 0.05$, $p_5 = 0.25$, and $p_6 = 0.07$. Suppose also that the die is to be rolled 40 times. Let X_1 denote the number of rolls for which an even number appears, and let X_2 denote the number of rolls for which either the number 1 or the number 3 appears. Find the value of $\Pr(X_1 = 20 \text{ and } X_2 = 15)$.

5. Suppose that 16 percent of the students in a certain high school are freshmen, 14 percent are sophomores, 38 percent are juniors, and 32 percent are seniors. If 15 students are selected at random from the school, what is the probability that at least eight will be either freshmen or sophomores?

6. In Exercise 5, let X_3 denote the number of juniors in the random sample of 15 students, and let X_4 denote the number of seniors in the sample. Find the value of $E(X_3 - X_4)$ and the value of $\text{Var}(X_3 - X_4)$.

7. Suppose that the random variables $X_1, \ldots, X_k$ are independent and that X_i has the Poisson distribution with mean λ_i $(i = 1, \ldots, k)$. Show that for each fixed positive integer n, the conditional distribution of the random vector $\boldsymbol{X} = (X_1, \ldots, X_k)$, given that $\sum_{i=1}^{k} X_i = n$, is the multinomial distribution with parameters n and $\boldsymbol{p} = (p_1, \ldots, p_k)$, where

$$p_i = \frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j} \quad \text{for } i = 1, \ldots, k.$$

8. Suppose that the parts produced by a machine can have three different levels of functionality: working, impaired, defective. Let p_1, p_2, and $p_3 = 1 - p_1 - p_2$ be the probabilities that a part is working, impaired, and defective, respectively. Suppose that the vector $\boldsymbol{p} = (p_1, p_2)$ is unknown but has a joint distribution with p.d.f.

$$f(p_1, p_2) = \begin{cases} 12p_1^2 & \text{for } 0 < p_1, p_2 < 1 \\ & \quad \text{and } p_1 + p_2 < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that we observe 10 parts that are conditionally independent given $\boldsymbol{p}$, and among those 10 parts, eight are working and two are impaired. Find the conditional p.d.f. of $\boldsymbol{p}$ given the observed parts. *Hint:* You might find Eq. (5.8.2) helpful.

5.10 The Bivariate Normal Distributions

The first family of multivariate continuous distributions for which we have a name is a generalization of the family of normal distributions to two coordinates. There is more structure to a bivariate normal distribution than just a pair of normal marginal distributions.

Definition and Derivation of Bivariate Normal Distributions

Example 5.10.1 Thyroid Hormones. Production of rocket fuel produces a chemical, perchlorate, that has found its way into drinking water supplies. Perchlorate is suspected of inhibiting thyroid function. Experiments have been performed in which laboratory rats have

been dosed with perchlorate in their drinking water. After several weeks, rats were sacrificed, and a number of thyroid hormones were measured. The levels of these hormones were then compared to the levels of the same hormones in rats that received no perchlorate in their water. Two hormones, TSH and T4, were of particular interest. Experimenters were interested in the joint distribution of TSH and T4. Although each of the hormones might be modeled with a normal distribution, a bivariate distribution is needed in order to model the two hormone levels jointly. Knowledge of thyroid activity suggests that the levels of these hormones will not be independent, because one of them is actually used by the thyroid to stimulate production of the other. ◀

If researchers are comfortable using the family of normal distributions to model each of two random variables separately, such as the hormones in Example 5.10.1, then they need a bivariate generalization of the family of normal distributions that still has normal distributions for its marginals while allowing the two random variables to be dependent. A simple way to create such a generalization is to make use of the result in Corollary 5.6.1. That result says that a linear combination of independent normal random variables has a normal distribution. If we create two different linear combinations X_1 and X_2 of the same independent normal random variables, then X_1 and X_2 will each have a normal distribution and they might be dependent. The following result formalizes this idea.

Theorem 5.10.1 Suppose that Z_1 and Z_2 are independent random variables, each of which has the standard normal distribution. Let $\mu_1, \mu_2, \sigma_1, \sigma_2$, and ρ be constants such that $-\infty < \mu_i < \infty$ $(i = 1, 2)$, $\sigma_i > 0$ $(i = 1, 2)$, and $-1 < \rho < 1$. Define two new random variables X_1 and X_2 as follows:

$$X_1 = \sigma_1 Z_1 + \mu_1,$$
$$X_2 = \sigma_2\left[\rho Z_1 + (1-\rho^2)^{1/2} Z_2\right] + \mu_2. \tag{5.10.1}$$

The joint p.d.f. of X_1 and X_2 is

$$f(x_1, x_2) = \frac{1}{2\pi(1-\rho^2)^{1/2}\sigma_1\sigma_2} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}. \tag{5.10.2}$$

Proof This proof relies on Theorem 3.9.5 (multivariate transformation of random variables). If you did not study Theorem 3.9.5, you won't be able to follow this proof. The joint p.d.f. $g(z_1, z_2)$ of Z_1 and Z_2 is

$$g(z_1, z_2) = \frac{1}{2\pi}\exp\left[-\frac{1}{2}(z_1^2 + z_2^2)\right], \tag{5.10.3}$$

for all z_1 and z_2.

The inverse of the transformation (5.10.1) is $(Z_1, Z_2) = (s_1(X_1, X_2), s_2(X_1, X_2))$, where

$$s_1(x_1, x_2) = \frac{x_1-\mu_1}{\sigma_1},$$
$$s_2(x_1, x_2) = \frac{1}{(1-\rho^2)^{1/2}}\left(\frac{x_2-\mu_2}{\sigma_2} - \rho\frac{x_1-\mu_1}{\sigma_1}\right). \tag{5.10.4}$$

The Jacobian J of the transformation is

$$J = \det \begin{bmatrix} \dfrac{1}{\sigma_1} & 0 \\ \dfrac{-\rho}{\sigma_1(1-\rho^2)^{1/2}} & \dfrac{1}{\sigma_2(1-\rho^2)^{1/2}} \end{bmatrix} = \frac{1}{(1-\rho^2)^{1/2}\sigma_1\sigma_2}. \qquad (5.10.5)$$

If one substitutes $s_i(x_1, x_2)$ for z_i $(i = 1, 2)$ in Eq. (5.10.3) and then multiplies by $|J|$, one obtains Eq. (5.10.2), which is the joint p.d.f. of (X_1, X_2) according to Theorem 3.9.5. ■

Some simple properties of the distribution with p.d.f. in Eq. (5.10.2) are worth deriving before giving a name to the joint distribution.

Theorem 5.10.2 Suppose that X_1 and X_2 have the joint distribution whose p.d.f. is given by Eq. (5.10.2). Then there exist independent standard normal random variables Z_1 and Z_2 such that Eqs. (5.10.1) hold. Also, the mean of X_i is μ_i and the variance of X_i is σ_i^2 for $i = 1, 2$. Furthermore the correlation between X_1 and X_2 is ρ. Finally, the marginal distribution of X_i is the normal distribution with mean μ_i and variance σ_i^2 for $i = 1, 2$.

Proof Use the functions s_1 and s_2 defined in Eqs. (5.10.4) and define $Z_i = s_i(X_1, X_2)$ for $i = 1, 2$. By running the proof of Theorem 5.10.1 in reverse, we see that the joint p.d.f. of Z_1 and Z_2 is Eq. (5.10.3). Hence, Z_1 and Z_2 are independent standard normal random variables.

The values of the means and variances of X_1 and X_2 are easily obtained by applying Corollary 5.6.1 to Eq. (5.10.1). If one applies the result in Exercise 8 of Sec. 4.6, one obtains $\text{Cov}(X_1, X_2) = \sigma_1\sigma_2\rho$. It now follows that ρ is the correlation. The claim about the marginal distributions of X_1 and X_2 is immediate from Corollary 5.6.1. ■

We are now ready to define the family of bivariate normal distributions.

Definition 5.10.1 Bivariate Normal Distributions. When the joint p.d.f. of two random variables X_1 and X_2 is of the form in Eq. (5.10.2), it is said that X_1 and X_2 have the *bivariate normal distribution with means* μ_1 *and* μ_2*, variances* σ_1^2 *and* σ_2^2*, and correlation* ρ.

It was convenient for us to derive the bivariate normal distributions as the joint distributions of certain linear combinations of independent random variables having standard normal distributions. It should be emphasized, however, that bivariate normal distributions arise directly and naturally in many practical problems. For example, for many populations the joint distribution of two physical characteristics such as the heights and the weights of the individuals in the population will be approximately a bivariate normal distribution. For other populations, the joint distribution of the scores of the individuals in the population on two related tests will be approximately a bivariate normal distribution.

Example 5.10.2 Anthropometry of Flea Beetles. Lubischew (1962) reports the measurements of several physical features of a variety of species of flea beetle. The investigation was concerned with whether some combination of easily obtained measurements could be used to distinguish the different species. Figure 5.9 shows a scatterplot of measurements of the first joint in the first tarsus versus the second joint in the first tarsus for a sample of 31 from the species *Chaetocnema heikertingeri*. The plot also includes three ellipses that correspond to a fitted bivariate normal distribution. The ellipses were chosen to contain 25%, 50%, and 75% of the probability of the fitted bivariate normal

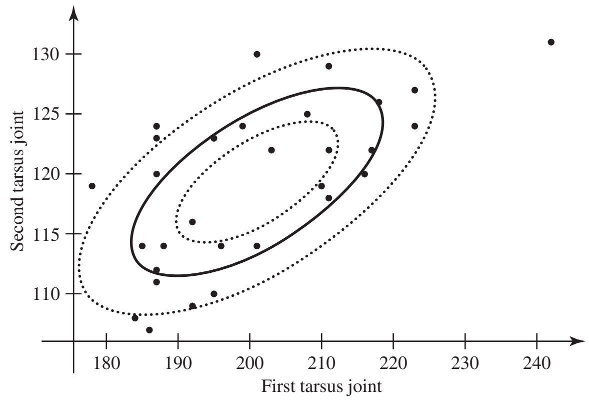


Figure 5.9 Scatterplot of flea beetle data with 25%, 50%, and 75% bivariate normal ellipses for Example 5.10.2.

distribution. The fitted distribution is is the bivariate normal distribution with means 201 and 119.3, variances 222.1 and 44.2, and correlation 0.64. ◀

Properties of Bivariate Normal Distributions

For random variables with a bivariate normal distribution, we find that being independent is equivalent to being uncorrelated.

Theorem 5.10.3 Independence and Correlation. Two random variables X_1 and X_2 that have a bivariate normal distribution are independent if and only if they are uncorrelated.

Proof The "only if" direction is already known from Theorem 4.6.4. For the "if" direction, assume that X_1 and X_2 are uncorrelated. Then $\rho = 0$, and it can be seen from Eq. (5.10.2) that the joint p.d.f. $f(x_1, x_2)$ factors into the product of the marginal p.d.f. of X_1 and the marginal p.d.f. of X_2. Hence, X_1 and X_2 are independent. ■

We have already seen in Example 4.6.4 that two random variables X_1 and X_2 with an arbitrary joint distribution can be uncorrelated without being independent. Theorem 5.10.3 says that no such examples exist in which X_1 and X_2 have a bivariate normal distribution.

When the correlation is not zero, Theorem 5.10.2 gives the marginal distributions of bivariate normal random variables. Combining the marginal and joint distributions allows us to find the conditional distributions of each X_i given the other one. The next theorem derives the conditional distributions using another technique.

Theorem 5.10.4 Conditional Distributions. Let X_1 and X_2 have the bivariate normal distribution whose p.d.f. is Eq. (5.10.2). The conditional distribution of X_2 given that $X_1 = x_1$ is the normal distribution with mean and variance given by

$$E(X_2|x_1) = \mu_2 + \rho\sigma_2\left(\frac{x_1 - \mu_1}{\sigma_1}\right), \quad \mathrm{Var}(X_2|x_1) = (1 - \rho^2)\sigma_2^2. \tag{5.10.6}$$

Proof We will make liberal use of Theorem 5.10.2 and its notation in this proof. Conditioning on $X_1 = x_1$ is the same as conditioning on $Z_1 = (x_1 - \mu_1)/\sigma_1$. When we want to find the conditional distribution of X_2 given $Z_1 = (x_1 - \mu_1)/\sigma_1$, we can subtitute $(x_1 - \mu_1)/\sigma_1$ for Z_2 in the formula for X_2 in Eq. (5.10.1) and find the conditional distribution for the rest of the formula. That is, the conditional distribution of X_2 given

that $X_1 = x_1$ is the same as the conditional distribution of

$$(1-\rho^2)^{1/2}\sigma_2 Z_2 + \mu_2 + \rho\sigma_2\left(\frac{x_1-\mu_1}{\sigma_1}\right) \tag{5.10.7}$$

given $Z_1 = (x_1 - \mu_1)/\sigma_1$. But Z_2 is the only random variable in Eq. (5.10.7), and Z_2 is independent of Z_1. Hence, the conditional distribution of X_2 given $X_1 = x_1$ is the marginal distribution of Eq. (5.10.7), namely, the normal distribution with mean and variance given by Eq. (5.10.6). ■

The conditional distribution of X_1 given that $X_2 = x_2$ cannot be derived so easily from Eq. (5.10.1) because of the different ways in which Z_1 and Z_2 enter Eq. (5.10.1). However, it is seen from Eq. (5.10.2) that the joint distribution of X_2 and X_1 is also bivariate normal with all of the subscripts 1 and 2 swithched on all of the parameters. Hence, we can apply Theorem 5.10.4 to X_2 and X_1 to conclude that the conditional distribution of X_1 given that $X_2 = x_2$ must be the normal distribution with mean and variance

$$E(X_1|x_2) = \mu_1 + \rho\sigma_1\left(\frac{x_2-\mu_2}{\sigma_2}\right), \quad \text{Var}(X_1|x_2) = (1-\rho^2)\sigma_1^2. \tag{5.10.8}$$

We have now shown that each marginal distribution and each conditional distribution of a bivariate normal distribution is a univariate normal distribution.

Some particular features of the conditional distribution of X_2 given that $X_1 = x_1$ should be noted. If $\rho \neq 0$, then $E(X_2|x_1)$ is a linear function of x_1. If $\rho > 0$, the slope of this linear function is positive. If $\rho < 0$, the slope of the function is negative. However, the variance of the conditional distribution of X_2 given that $X_1 = x_1$ is $(1-\rho^2)\sigma_2^2$, which does not depend on x_1. Furthermore, this variance of the conditional distribution of X_2 is smaller than the variance σ_2^2 of the marginal distribution of X_2.

Example 5.10.3

Predicting a Person's Weight. Let X_1 denote the height of a person selected at random from a certain population, and let X_2 denote the weight of the person. Suppose that these random variables have the bivariate normal distribution for which the p.d.f. is specified by Eq. (5.10.2) and that the person's weight X_2 must be predicted. We shall compare the smallest M.S.E. that can be attained if the person's height X_1 is known when her weight must be predicted with the smallest M.S.E. that can be attained if her height is not known.

If the person's height is not known, then the best prediction of her weight is the mean $E(X_2) = \mu_2$, and the M.S.E. of this prediction is the variance σ_2^2. If it is known that the person's height is x_1, then the best prediction is the mean $E(X_2|x_1)$ of the conditional distribution of X_2 given that $X_1 = x_1$, and the M.S.E. of this prediction is the variance $(1-\rho^2)\sigma_2^2$ of that conditional distribution. Hence, when the value of X_1 is known, the M.S.E. is reduced from σ_2^2 to $(1-\rho^2)\sigma_2^2$. ◂

Since the variance of the conditional distribution in Example 5.10.3 is $(1-\rho^2)\sigma_2^2$, regardless of the known height x_1 of the person, it follows that the difficulty of predicting the person's weight is the same for a tall person, a short person, or a person of medium height. Furthermore, since the variance $(1-\rho^2)\sigma_2^2$ decreases as $|\rho|$ increases, it follows that it is easier to predict a person's weight from her height when the person is selected from a population in which height and weight are highly correlated.

Example 5.10.4 **Determining a Marginal Distribution.** Suppose that a random variable X has the normal distribution with mean μ and variance σ^2, and that for every number x, the conditional distribution of another random variable Y given that $X = x$ is the normal distribution with mean x and variance τ^2. We shall determine the marginal distribution of Y.

We know that the marginal distribution of X is a normal distribution, and the conditional distribution of Y given that $X = x$ is a normal distribution, for which the mean is a linear function of x and the variance is constant. It follows that the joint distribution of X and Y must be a bivariate normal distribution (see Exercise 14). Hence, the marginal distribution of Y is also a normal distribution. The mean and the variance of Y must be determined.

The mean of Y is

$$E(Y) = E[E(Y|X)] = E(X) = \mu.$$

Furthermore, by Theorem 4.7.4,

$$\begin{aligned}\mathrm{Var}(Y) &= E[\mathrm{Var}(Y|X)] + \mathrm{Var}[E(Y|X)]\\ &= E(\tau^2) + \mathrm{Var}(X)\\ &= \tau^2 + \sigma^2.\end{aligned}$$

Hence, the distribution of Y is the normal distribution with mean μ and variance $\tau^2 + \sigma^2$. ◄

Linear Combinations

Example 5.10.5 **Heights of Husbands and Wives.** Suppose that a married couple is selected at random from a certain population of married couples and that the joint distribution of the height of the wife and the height of her husband is a bivariate normal distribution. What is the probability that, in the randomly chosen couple, the wife is taller than the husband? ◄

The question asked at the end of Example 5.10.5 can be expressed in terms of the distribution of the difference between a wife's and husband's heights. This is a special case of a linear combination of a bivariate normal vector.

Theorem 5.10.5 **Linear Combination of Bivariate Normals.** Suppose that two random variables X_1 and X_2 have a bivariate normal distribution, for which the p.d.f. is specified by Eq. (5.10.2). Let $Y = a_1X_1 + a_2X_2 + b$, where a_1, a_2, and b are arbitrary given constants. Then Y has the normal distribution with mean $a_1\mu_1 + a_2\mu_2 + b$ and variance

$$a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2\rho\sigma_1\sigma_2. \tag{5.10.9}$$

Proof According to Theorem 5.10.2, both X_1 and X_2 can be represented, as in Eq. (5.10.1), as linear combinations of independent and normally distributed random variables Z_1 and Z_2. Since Y is a linear combination of X_1 and X_2, it follows that Y can also be represented as a linear combination of Z_1 and Z_2. Therefore, by Corollary 5.6.1, the distribution of Y will also be a normal distribution. It only remains to compute the mean and variance of Y. The mean of Y is

$$\begin{aligned}E(Y) &= a_1E(X_1) + a_2E(X_2) + b\\ &= a_1\mu_1 + a_2\mu_2 + b.\end{aligned}$$

It also follows from Corollary 4.6.1 that

$$\text{Var}(Y) = a_1^2 \,\text{Var}(X_1) + a_2^2 \,\text{Var}(X_2) + 2a_1 a_2 \,\text{Cov}(X_1, X_2).$$

That $\text{Var}(Y)$ is given by Eq. (5.10.9) now follows easily. ■

Example 5.10.6

Heights of Husbands and Wives. Consider again Example 5.10.5. Suppose that the heights of the wives have a mean of 66.8 inches and a standard deviation of 2 inches, the heights of the husbands have a mean of 70 inches and a standard deviation of 2 inches, and the correlation between these two heights is 0.68. We shall determine the probability that the wife will be taller than her husband.

If we let X denote the height of the wife, and let Y denote the height of her husband, then we must determine the value of $\Pr(X - Y > 0)$. Since X and Y have a bivariate normal distribution, it follows that the distribution of $X - Y$ will be the normal distribution, with mean

$$E(X - Y) = 66.8 - 70 = -3.2$$

and variance

$$\begin{aligned}\text{Var}(X - Y) &= \text{Var}(X) + \text{Var}(Y) - 2\,\text{Cov}(X, Y)\\ &= 4 + 4 - 2(0.68)(2)(2) = 2.56.\end{aligned}$$

Hence, the standard deviation of $X - Y$ is 1.6.

The random variable $Z = (X - Y + 3.2)/(1.6)$ will have the standard normal distribution. It can be found from the table given at the end of this book that

$$\begin{aligned}\Pr(X - Y > 0) = \Pr(Z > 2) &= 1 - \Phi(2)\\ &= 0.0227.\end{aligned}$$

Therefore, the probability that the wife will be taller than her husband is 0.0227. ◄

Summary

If a random vector (X, Y) has a bivariate normal distribution, then every linear combination $aX + bY + c$ has a normal distribution. In particular, the marginal distributions of X and Y are normal. Also, the conditional distribution of X given $Y = y$ is normal with the conditional mean being a linear function of y and the conditional variance being constant in y. (Similarly, for the conditional distribution of Y given $X = x$.) A more thorough treatment of the bivariate normal distributions and higher-dimensional generalizations can be found in the book by D. F. Morrison (1990).

Exercises

1. Consider again the joint distribution of heights of husbands and wives in Example 5.10.6. Find the 0.95 quantile of the conditional distribution of the height of the wife given that the height of the husband is 72 inches.

2. Suppose that two different tests A and B are to be given to a student chosen at random from a certain population. Suppose also that the mean score on test A is 85, and the standard deviation is 10; the mean score on test B is 90, and the standard deviation is 16; the scores on the two tests have a bivariate normal distribution; and the correlation of the two scores is 0.8. If the student's score on test A is 80, what is the probability that her score on test B will be higher than 90?

3. Consider again the two tests A and B described in Exercise 2. If a student is chosen at random, what is the probability that the sum of her scores on the two tests will be greater than 200?

4. Consider again the two tests A and B described in Exercise 2. If a student is chosen at random, what is the probability that her score on test A will be higher than her score on test B?

5. Consider again the two tests A and B described in Exercise 2. If a student is chosen at random, and her score on test B is 100, what predicted value of her score on test A has the smallest M.S.E., and what is the value of this minimum M.S.E.?

6. Suppose that the random variables X_1 and X_2 have a bivariate normal distribution, for which the joint p.d.f. is specified by Eq. (5.10.2). Determine the value of the constant b for which $\text{Var}(X_1 + bX_2)$ will be a minimum.

7. Suppose that X_1 and X_2 have a bivariate normal distribution for which $E(X_1|X_2) = 3.7 - 0.15X_2$, $E(X_2|X_1) = 0.4 - 0.6X_1$, and $\text{Var}(X_2|X_1) = 3.64$. Find the mean and the variance of X_1, the mean and the variance of X_2, and the correlation of X_1 and X_2.

8. Let $f(x_1, x_2)$ denote the p.d.f. of the bivariate normal distribution specified by Eq. (5.10.2). Show that the maximum value of $f(x_1, x_2)$ is attained at the point at which $x_1 = \mu_1$ and $x_2 = \mu_2$.

9. Let $f(x_1, x_2)$ denote the p.d.f. of the bivariate normal distribution specified by Eq. (5.10.2), and let k be a constant such that

$$0 < k < \frac{1}{2\pi(1-\rho^2)^{1/2}\sigma_1\sigma_2}.$$

Show that the points (x_1, x_2) such that $f(x_1, x_2) = k$ lie on a circle if $\rho = 0$ and $\sigma_1 = \sigma_2$, and these points lie on an ellipse otherwise.

10. Suppose that two random variables X_1 and X_2 have a bivariate normal distribution, and two other random variables Y_1 and Y_2 are defined as follows:

$$Y_1 = a_{11}X_1 + a_{12}X_2 + b_1,$$
$$Y_2 = a_{21}X_1 + a_{22}X_2 + b_2,$$

where

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0.$$

Show that Y_1 and Y_2 also have a bivariate normal distribution.

11. Suppose that two random variables X_1 and X_2 have a bivariate normal distribution, and $\text{Var}(X_1) = \text{Var}(X_2)$. Show that the sum $X_1 + X_2$ and the difference $X_1 - X_2$ are independent random variables.

12. Suppose that the two measurements from flea beetles in Example 5.10.2 have the bivariate normal distribution with $\mu_1 = 201$, $\mu_2 = 118$, $\sigma_1 = 15.2$, $\sigma_2 = 6.6$, and $\rho = 0.64$. Suppose that the same two measurements from a second species also have the bivariate normal distribution with $\mu_1 = 187$, $\mu_2 = 131$, $\sigma_1 = 15.2$, $\sigma_2 = 6.6$, and $\rho = 0.64$. Let (X_1, X_2) be a pair of measurements on a flea beetle from one of these two species. Let a_1, a_2 be constants.

- **a.** For each of the two species, find the mean and standard deviation of $a_1X_1 + a_2X_2$. (Note that the variances for the two species will be the same. How do you know that?)
- **b.** Find a_1 and a_2 to maximize the ratio of the difference between the two means found in part (a) to the standard deviation found in part (a). There is a sense in which this linear combination $a_1X_1 + a_2X_2$ does the best job of distinguishing the two species among all possible linear combinations.

13. Suppose that the joint p.d.f. of two random variables X and Y is proportional, as a function of (x, y), to

$$\exp\Big(-[ax^2 + by^2 + cxy + ex + gy + h]\Big),$$

where $a > 0$, $b > 0$, and c, e, g, and h are all constants. Assume that $ab > (c/2)^2$. Prove that X and Y have a bivariate normal distribution, and find the means, variances, and correlation.

14. Suppose that a random variable X has a normal distribution, and for every x, the conditional distribution of another random variable Y given that $X = x$ is a normal distribution with mean $ax + b$ and variance τ^2, where a, b, and τ^2 are constants. Prove that the joint distribution of X and Y is a bivariate normal distribution.

15. Let $X_1, \ldots, X_n$ be i.i.d. random variables having the normal distribution with mean μ and variance σ^2. Define $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, the sample mean. In this problem, we shall find the conditional distribution of each X_i given $\overline{X}_n$.

- **a.** Show that X_i and $\overline{X}_n$ have the bivariate normal distribution with both means μ, variances σ^2 and σ^2/n, and correlation $1/\sqrt{n}$. *Hint:* Let $Y = \sum_{j \neq i} X_j$. Now show that Y and X_i are independent normals and $\overline{X}_n$ and X_i are linear combinations of Y and X_i.
- **b.** Show that the conditional distribution of X_i given $\overline{X}_n = \overline{x}_n$ is normal with mean $\overline{x}_n$ and variance $\sigma^2(1 - 1/n)$.

5.11 Supplementary Exercises

1. Let X and P be random variables. Suppose that the conditional distribution of X given $P = p$ is the binomial distribution with parameters n and p. Suppose that the distribution of P is the beta distribution with parameters $\alpha = 1$ and $\beta = 1$. Find the marginal distribution of X.

2. Suppose that X, Y, and Z are i.i.d. random variables and each has the standard normal distribution. Evaluate $\Pr(3X + 2Y < 6Z - 7)$.

3. Suppose that X and Y are independent Poisson random variables such that $\text{Var}(X) + \text{Var}(Y) = 5$. Evaluate $\Pr(X + Y < 2)$.

4. Suppose that X has a normal distribution such that $\Pr(X < 116) = 0.20$ and $\Pr(X < 328) = 0.90$. Determine the mean and the variance of X.

5. Suppose that a random sample of four observations is drawn from the Poisson distribution with mean λ, and let $\overline{X}$ denote the sample mean. Show that

$$\Pr\left(\overline{X} < \frac{1}{2}\right) = (4\lambda + 1)e^{-4\lambda}.$$

6. The lifetime X of an electronic component has the exponential distribution such that $\Pr(X \le 1000) = 0.75$. What is the expected lifetime of the component?

7. Suppose that X has the normal distribution with mean μ and variance σ^2. Express $E(X^3)$ in terms of μ and σ^2.

8. Suppose that a random sample of 16 observations is drawn from the normal distribution with mean μ and standard deviation 12, and that independently another random sample of 25 observations is drawn from the normal distribution with the same mean μ and standard deviation 20. Let $\overline{X}$ and $\overline{Y}$ denote the sample means of the two samples. Evaluate $\Pr(|\overline{X} - \overline{Y}| < 5)$.

9. Suppose that men arrive at a ticket counter according to a Poisson process at the rate of 120 per hour, and women arrive according to an independent Poisson process at the rate of 60 per hour. Determine the probability that four or fewer people arrive in a one-minute period.

10. Suppose that $X_1, X_2, \ldots$ are i.i.d. random variables, each of which has m.g.f. $\psi(t)$. Let $Y = X_1 + \cdots + X_N$, where the number of terms N in this sum is a random variable having the Poisson distribution with mean λ. Assume that N and $X_1, X_2, \ldots$ are independent, and $Y = 0$ if $N = 0$. Determine the m.g.f. of Y.

11. Every Sunday morning, two children, Craig and Jill, independently try to launch their model airplanes. On each Sunday, Craig has probability 1/3 of a successful launch, and Jill has probability 1/5 of a successful launch. Determine the expected number of Sundays required until at least one of the two children has a successful launch.

12. Suppose that a fair coin is tossed until at least one head and at least one tail have been obtained. Let X denote the number of tosses that are required. Find the p.f. of X.

13. Suppose that a pair of balanced dice are rolled 120 times, and let X denote the number of rolls on which the sum of the two numbers is 12. Use the Poisson approximation to approximate $\Pr(X = 3)$.

14. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval [0, 1]. Let $Y_1 = \min\{X_1, \ldots, X_n\}$, $Y_n = \max\{X_1, \ldots, X_n\}$, and $W = Y_n - Y_1$. Show that each of the random variables Y_1, Y_n, and W has a beta distribution.

15. Suppose that events occur in accordance with a Poisson process at the rate of five events per hour.

- **a.** Determine the distribution of the waiting time T_1 until the first event occurs.
- **b.** Determine the distribution of the total waiting time T_k until k events have occurred.
- **c.** Determine the probability that none of the first k events will occur within 20 minutes of one another.

16. Suppose that five components are functioning simultaneously, that the lifetimes of the components are i.i.d., and that each lifetime has the exponential distribution with parameter β. Let T_1 denote the time from the beginning of the process until one of the components fails; and let T_5 denote the total time until all five components have failed. Evaluate $\text{Cov}(T_1, T_5)$.

17. Suppose that X_1 and X_2 are independent random variables, and X_i has the exponential distribution with parameter β_i $(i = 1, 2)$. Show that for each constant $k > 0$,

$$\Pr(X_1 > kX_2) = \frac{\beta_2}{k\beta_1 + \beta_2}.$$

18. Suppose that 15,000 people in a city with a population of 500,000 are watching a certain television program. If 200 people in the city are contacted at random, what is the approximate probability that fewer than four of them are watching the program?

19. Suppose that it is desired to estimate the proportion of persons in a large population who have a certain characteristic. A random sample of 100 persons is selected from the population without replacement, and the proportion $\overline{X}$ of persons in the sample who have the characteristic is observed. Show that, no matter how large the population is, the standard deviation of $\overline{X}$ is at most 0.05.

20. Suppose that X has the binomial distribution with parameters n and p, and that Y has the negative binomial distribution with parameters r and p, where r is a positive integer. Show that $\Pr(X < r) = \Pr(Y > n - r)$ by showing

that both the left side and the right side of this equation can be regarded as the probability of the same event in a sequence of Bernoulli trials with probability p of success.

21. Suppose that X has the Poisson distribution with mean λt, and that Y has the gamma distribution with parameters $\alpha = k$ and $\beta = \lambda$, where k is a positive integer. Show that $\Pr(X \geq k) = \Pr(Y \leq t)$ by showing that both the left side and the right side of this equation can be regarded as the probability of the same event in a Poisson process in which the expected number of occurrences per unit of time is λ.

22. Suppose that X is a random variable having a continuous distribution with p.d.f. $f(x)$ and c.d.f. $F(x)$, and for which $\Pr(X > 0) = 1$. Let the failure rate $h(x)$ be as defined in Exercise 18 of Sec. 5.7. Show that

$$\exp\left[-\int_0^x h(t)\,dt\right] = 1 - F(x).$$

23. Suppose that 40 percent of the students in a large population are freshmen, 30 percent are sophomores, 20 percent are juniors, and 10 percent are seniors. Suppose that 10 students are selected at random from the population, and let X_1, X_2, X_3, X_4 denote, respectively, the numbers of freshmen, sophomores, juniors, and seniors that are obtained.

a. Determine $\rho(X_i, X_j)$ for each pair of values i and j $(i < j)$.

b. For what values of i and j $(i < j)$ is $\rho(X_i, X_j)$ most negative?

c. For what values of i and j $(i < j)$ is $\rho(X_i, X_j)$ closest to 0?

24. Suppose that X_1 and X_2 have the bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ. Determine the distribution of $X_1 - 3X_2$.

25. Suppose that X has the standard normal distribution, and the conditional distribution of Y given X is the normal distribution with mean $2X - 3$ and variance 12. Determine the marginal distribution of Y and the value of $\rho(X, Y)$.

26. Suppose that X_1 and X_2 have a bivariate normal distribution with $E(X_2) = 0$. Evaluate $E(X_1^2 X_2)$.

Chapter 6

LARGE RANDOM SAMPLES

6.1 Introduction

In this chapter, we introduce a number of approximation results that simplify the analysis of large random samples. In the first section, we give two examples to illustrate the types of analyses that we might wish to perform and how additional tools may be needed to be able to perform them.

Example 6.1.1

Proportion of Heads. If you draw a coin from your pocket, you might feel confident that it is essentially fair. That is, the probability that it will land with head up when flipped is 1/2. However, if you were to flip the coin 10 times, you would not expect to see exactly 5 heads. If you were to flip it 100 times, you would be even less likely to see exactly 50 heads. Indeed, we can calculate the probabilities of each of these two results using the fact that the number of heads in n independent flips of a fair coin has the binomial distribution with parameters n and 1/2. So, if X is the number of heads in 10 independent flips, we know that

$$\Pr(X = 5) = \binom{10}{5}\left(\frac{1}{2}\right)^5\left(1-\frac{1}{2}\right)^5 = 0.2461.$$

If Y is the number of heads in 100 independent flips, we have

$$\Pr(Y = 50) = \binom{100}{50}\left(\frac{1}{2}\right)^{50}\left(1-\frac{1}{2}\right)^{50} = 0.0796.$$

Even though the probability of exactly $n/2$ heads in n flips is quite small, especially for large n, you still expect the proportion of heads to be close to 1/2 if n is large. For example, if $n = 100$, the proportion of heads is $Y/100$. In this case, the probability that the proportion is within 0.1 of 1/2 is

$$\Pr\left(0.4 \le \frac{Y}{100} \le 0.6\right) = \Pr(40 \le Y \le 60) = \sum_{i=40}^{60}\binom{100}{i}\left(\frac{1}{2}\right)^i\left(1-\frac{1}{2}\right)^{100-i} = 0.9648.$$

A similar calculation with $n = 10$ yields

$$\Pr\left(0.4 \le \frac{X}{10} \le 0.6\right) = \Pr(4 \le Y \le 6) = \sum_{i=4}^{6}\binom{10}{i}\left(\frac{1}{2}\right)^i\left(1-\frac{1}{2}\right)^{10-i} = 0.6563.$$

Notice that the probability that the proportion of heads in n tosses is close to 1/2 is larger for $n = 100$ than for $n = 10$ in this example. This is due in part to the fact that

we have defined "close to 1/2" to be the same for both cases, namely, between 0.4 and 0.6. ◄

The calculations performed in Example 6.1.1 were simple enough because we have a formula for the probability function of the number of heads in any number of flips. For more complicated random variables, the situation is not so simple.

Example 6.1.2 Average Waiting Time. A queue is serving customers, and the ith customer waits a random time X_i to be served. Suppose that $X_1, X_2, \dots$ are i.i.d. random variables having the uniform distribution on the interval $[0, 1]$. The mean waiting time is 0.5. Intuition suggests that the average of a large number of waiting times should be close to the mean waiting time. But the distribution of the average of $X_1, \dots, X_n$ is rather complicated for every $n > 1$. It may not be possible to calculate precisely the probability that the sample average is close to 0.5 for large samples. ◄

The law of large numbers (Theorem 6.2.4) will give a mathematical foundation to the intuition that the average of a large sample of i.i.d. random variables, such as the waiting times in Example 6.1.2, should be close to their mean. The central limit theorem (Theorem 6.3.1) will give us a way to approximate the probability that the sample average is close to the mean.

Exercises

1. The solution to Exercise 1 of Sec. 3.9 is the p.d.f. of $X_1 + X_2$ in Example 6.1.2. Find the p.d.f. of $\overline{X}_2 = (X_1 + X_2)/2$. Compare the probabilities that $\overline{X}_2$ and X_1 are close to 0.5. In particular, compute $\Pr(|\overline{X}_2 - 0.5| < 0.1)$ and $\Pr(|X_1 - 0.5| < 0.1)$. What feature of the p.d.f. of $\overline{X}_2$ makes it clear that the distribution is more concentrated near the mean?

2. Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables having the normal distribution with mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean of the first n random variables in the sequence. Show that $\Pr(|\overline{X}_n - \mu| \le c)$ converges to 1 as $n \to \infty$. *Hint:* Write the probability in terms of the standard normal c.d.f. Φ and use what you know about this c.d.f.

3. This problem requires a computer program because the calculation is too tedious to do by hand. Extend the calculation in Example 6.1.1 to the case of $n = 200$ flips. That is, let W be the number of heads in 200 flips of a fair coin, and compute $\Pr\left(0.4 \le \frac{W}{200} \le 0.6\right)$. What do you think is the continuation of the pattern of these probabilities as the number of flips n increases without bound?

6.2 The Law of Large Numbers

The average of a random sample of i.i.d. random variables is called their sample mean. The sample mean is useful for summarizing the information in a random sample in much the same way that the mean of a probability distribution summarizes the information in the distribution. In this section, we present some results that illustrate the connection between the sample mean and the expected value of the individual random variables that comprise the random sample.

The Markov and Chebyshev Inequalities

We shall begin this section by presenting two simple and general results, known as the Markov inequality and the Chebyshev inequality. We shall then apply these inequalities to random samples.

The Markov inequality is related to the claim made on page 211 about how the mean of a distribution can be affected by moving a small amount of probability to an arbitrarily large value. The Markov inequality puts a bound on how much probability can be at arbitrarily large values once the mean is specified.

Theorem 6.2.1 Markov Inequality. Suppose that X is a random variable such that $\Pr(X \geq 0) = 1$. Then for every real number $t > 0$,

$$\Pr(X \geq t) \leq \frac{E(X)}{t}. \tag{6.2.1}$$

Proof For convenience, we shall assume that X has a discrete distribution for which the p.f. is f. The proof for a continuous distribution or a more general type of distribution is similar. For a discrete distribution,

$$E(X) = \sum_{x} xf(x) = \sum_{x<t} xf(x) + \sum_{x \geq t} xf(x).$$

Since X can have only nonnegative values, all the terms in the summations are nonnegative. Therefore,

$$E(X) \geq \sum_{x \geq t} xf(x) \geq \sum_{x \geq t} tf(x) = t \Pr(X \geq t). \tag{6.2.2}$$

Divide the extreme ends of (6.2.2) by $t > 0$ to obtain (6.2.1). ■

The Markov inequality is primarily of interest for large values of t. In fact, when $t \leq E(X)$, the inequality is of no interest whatsoever, since it is known that $\Pr(X \leq t) \leq 1$. However, it is found from the Markov inequality that for every nonnegative random variable X whose mean is 1, the maximum possible value of $\Pr(X \geq 100)$ is 0.01. Furthermore, it can be verified that this maximum value is actually attained by every random variable X for which $\Pr(X = 0) = 0.99$ and $\Pr(X = 100) = 0.01$.

The Chebyshev inequality is related to the idea that the variance of a random variable is a measure of how spread out its distribution is. The inequality says that the probability that X is far away from its mean is bounded by a quantity that increases as $\text{Var}(X)$ increases.

Theorem 6.2.2 Chebyshev Inequality. Let X be a random variable for which $\text{Var}(X)$ exists. Then for every number $t > 0$,

$$\Pr(|X - E(X)| \geq t) \leq \frac{\text{Var}(X)}{t^2}. \tag{6.2.3}$$

Proof Let $Y = [X - E(X)]^2$. Then $\Pr(Y \geq 0) = 1$ and $E(Y) = \text{Var}(X)$. By applying the Markov inequality to Y, we obtain the following result:

$$\Pr(|X - E(X)| \geq t) = \Pr(Y \geq t^2) \leq \frac{\text{Var}(X)}{t^2}.$$ ■

It can be seen from this proof that the Chebyshev inequality is simply a special case of the Markov inequality. Therefore, the comments that were given following the proof of the Markov inequality can be applied as well to the Chebyshev inequality. Because of their generality, these inequalities are very useful. For example, if $\text{Var}(X) = \sigma^2$ and we let $t = 3\sigma$, then the Chebyshev inequality yields the result that

$$\Pr(|X - E(X)| \geq 3\sigma) \leq \frac{1}{9}.$$

In words, the probability that any given random variable will differ from its mean by more than 3 standard deviations *cannot* exceed 1/9. This probability will actually be much smaller than 1/9 for many of the random variables and distributions that will be discussed in this book. The Chebyshev inequality is useful because of the fact that this probability must be 1/9 or less for *every* distribution. It can also be shown (see Exercise 4 at the end of this section) that the upper bound in (6.2.3) is sharp in the sense that it cannot be made any smaller and still hold for *all* distributions.

Properties of the Sample Mean

In Definition 5.6.3, we defined the *sample mean* of n random variables $X_1, \dots, X_n$ to be their average,

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n).$$

The mean and the variance of $\overline{X}_n$ are easily computed.

Theorem 6.2.3 **Mean and Variance of the Sample Mean.** Let $X_1, \dots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Let $\overline{X}_n$ be the sample mean. Then $E(\overline{X}_n) = \mu$ and $\text{Var}(\overline{X}_n) = \sigma^2/n$.

Proof It follows from Theorems 4.2.1 and 4.2.4 that

$$E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} \cdot n\mu = \mu.$$

Furthermore, since $X_1, \dots, X_n$ are independent, Theorems 4.3.4 and 4.3.5 say that

$$\begin{aligned}\text{Var}(\overline{X}_n) &= \frac{1}{n^2}\,\text{Var}\left(\sum_{i=1}^{n} X_i\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

■

In words, the mean of $\overline{X}_n$ is equal to the mean of the distribution from which the random sample was drawn, but the variance of $\overline{X}_n$ is only $1/n$ times the variance of that distribution. It follows that the probability distribution of $\overline{X}_n$ will be more concentrated around the mean value μ than was the original distribution. In other words, the sample mean $\overline{X}_n$ is more likely to be close to μ than is the value of just a single observation X_i from the given distribution.

These statements can be made more precise by applying the Chebyshev inequality to $\overline{X}_n$. Since $E(\overline{X}_n) = \mu$ and $\text{Var}(\overline{X}_n) = \sigma^2/n$, it follows from the relation (6.2.3) that for every number $t > 0$,

$$\Pr(|\overline{X}_n - \mu| \ge t) \le \frac{\sigma^2}{nt^2}. \tag{6.2.4}$$

Example 6.2.1 **Determining the Required Number of Observations.** Suppose that a random sample is to be taken from a distribution for which the value of the mean μ is not known, but for which it is known that the standard deviation σ is 2 units or less. We shall determine how large the sample size must be in order to make the probability at least 0.99 that $|\overline{X}_n - \mu|$ will be less than 1 unit.

Since $\sigma^2 \le 2^2 = 4$, it follows from the relation (6.2.4) that for every sample size n,

$$\Pr(|\overline{X}_n - \mu| \ge 1) \le \frac{\sigma^2}{n} \le \frac{4}{n}.$$

Since n must be chosen so that $\Pr(|\overline{X}_n - \mu| < 1) \ge 0.99$, it follows that n must be chosen so that $4/n \le 0.01$. Hence, it is required that $n \ge 400$. ◀

Example 6.2.2

A Simulation. An environmental engineer believes that there are two contaminants in a water supply, arsenic and lead. The actual concentrations of the two contaminants are independent random variables X and Y, measured in the same units. The engineer is interested in what proportion of the contamination is lead on average. That is, the engineer wants to know the mean of $R = Y/(X + Y)$. We suppose that it is a simple matter to generate as many independent pseudo-random numbers with the distributions of X and Y as we desire. A common way to obtain an approximation to $E[Y/(X + Y)]$ would be the following: If we sample n pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ and compute $R_i = Y_i/(X_i + Y_i)$ for $i = 1, \dots, n$, then $\overline{R}_n = \frac{1}{n}\sum_{i=1}^n R_i$ is a sensible approximation to $E(R)$. To decide how large n should be, we can argue as in Example 6.2.1. Since it is known that $|R_i| \le 1$, it must be that $\text{Var}(R_i) \le 1$. (Actually, $\text{Var}(R_i) \le 1/4$, but this is harder to prove. See Exercise 14 in this section for a way to prove it in the discrete case.) According to Chebyshev's inequality, for each $\epsilon > 0$,

$$\Pr\Big(|\overline{R}_n - E(R)| \ge \epsilon\Big) \le \frac{1}{n\epsilon^2}.$$

So, if we want $|\overline{R}_n - E(R)| \le 0.005$ with probability 0.98 or more, then we should use $n > 1/[0.2 \times 0.005^2] = 2{,}000{,}000$. ◀

It should be emphasized that the use of the Chebyshev inequality in Example 6.2.1 guarantees that a sample for which $n = 400$ will be large enough to meet the specified probability requirements, regardless of the particular type of distribution from which the sample is to be taken. If further information about this distribution is available, then it can often be shown that a smaller value for n will be sufficient. This property is illustrated in the next example.

Example 6.2.3

Tossing a Coin. Suppose that a fair coin is to be tossed n times independently. For $i = 1, \dots, n$, let $X_i = 1$ if a head is obtained on the ith toss, and let $X_i = 0$ if a tail is obtained on the ith toss. Then the sample mean $\overline{X}_n$ will simply be equal to the proportion of heads that are obtained on the n tosses. We shall determine the number of times the coin must be tossed in order to make $\Pr(0.4 \le \overline{X}_n \le 0.6) \ge 0.7$. We shall determine this number in two ways: first, by using the Chebyshev inequality; second, by using the exact probabilities for the binomial distribution of the total number of heads.

Let $T = \sum_{i=1}^n X_i$ denote the total number of heads that are obtained when n tosses are made. Then T has the binomial distribution with parameters n and $p = 1/2$. Therefore, it follows from Eq. (4.2.5) on page 221 that $E(T) = n/2$, and it follows from Eq. (4.3.3) on page 232 that $\text{Var}(T) = n/4$. Because $\overline{X}_n = T/n$, we can obtain

the following relation from the Chebyshev inequality:

$$\begin{aligned}\Pr(0.4 \le \overline{X}_n \le 0.6) &= \Pr(0.4n \le T \le 0.6n)\\ &= \Pr\left(\left|T - \frac{n}{2}\right| \le 0.1n\right)\\ &\ge 1 - \frac{n}{4(0.1n)^2} = 1 - \frac{25}{n}.\end{aligned}$$

Hence, if $n \ge 84$, this probability will be at least 0.7, as required.

However, from the table of binomial distributions given at the end of this book, it is found that for $n = 15$,

$$\Pr(0.4 \le \overline{X}_n \le 0.6) = \Pr(6 \le T \le 9) = 0.70.$$

Hence, 15 tosses would actually be sufficient to satisfy the specified probability requirement. ◀

The Law of Large Numbers

The discussion in Example 6.2.3 indicates that the Chebyshev inequality may not be a practical tool for determining the appropriate sample size in a particular problem, because it may specify a much greater sample size than is actually needed for the particular distribution from which the sample is being taken. However, the Chebyshev inequality is a valuable theoretical tool, and it will be used here to prove an important result known as the *law of large numbers*.

Suppose that $Z_1, Z_2, \ldots$ is a sequence of random variables. Roughly speaking, it is said that this sequence converges to a given number b if the probability distribution of Z_n becomes more and more concentrated around b as $n \to \infty$. To be more precise, we give the following definition.

Definition 6.2.1 Convergence in Probability. A sequence $Z_1, Z_2, \ldots$ of random variables *converges to b in probability* if for every number $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(|Z_n - b| < \varepsilon) = 1.$$

This property is denoted by

$$Z_n \xrightarrow{p} b,$$

and is sometimes stated simply as Z_n converges to b in probability.

In other words, Z_n converges to b in probability if the probability that Z_n lies in each given interval around b, no matter how small this interval may be, approaches 1 as $n \to \infty$.

We shall now show that the sample mean of a random sample with finite variance always converges in probability to the mean of the distribution from which the random sample was taken.

Theorem 6.2.4 Law of Large Numbers. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the mean is μ and for which the variance is finite. Let $\overline{X}_n$ denote the sample mean. Then

$$\overline{X}_n \xrightarrow{p} \mu. \tag{6.2.5}$$

Proof Let the variance of each X_i be σ^2. It then follows from the Chebyshev inequality that for every number $\varepsilon > 0$,

$$\Pr(|\overline{X}_n - \mu| < \varepsilon) \ge 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

Hence,

$$\lim_{n\to\infty} \Pr(|\overline{X}_n - \mu| < \varepsilon) = 1,$$

which means that $\overline{X}_n \xrightarrow{p} \mu$. ■

It can also be shown that Eq. (6.2.5) is satisfied if the distribution from which the random sample is taken has a finite mean μ but an infinite variance. However, the proof for this case is beyond the scope of this book.

Since $\overline{X}_n$ converges to μ in probability, it follows that there is high probability that $\overline{X}_n$ will be close to μ if the sample size n is large. Hence, if a large random sample is taken from a distribution for which the mean is unknown, then the arithmetic average of the values in the sample will usually be a close estimate of the unknown mean. This topic will be discussed again in Sec. 6.3, where we introduce the central limit theorem. It will then be possible to present a more precise probability distribution for the difference between $\overline{X}_n$ and μ.

The following result can be useful if we observe random variables with mean μ but are interested in μ^2 or $\log(\mu)$ or some other continuous function of μ. The proof is left for the reader (Exercise 15).

Theorem 6.2.5 Continuous Functions of Random Variables. If $Z_n \xrightarrow{p} b$, and if $g(z)$ is a function that is continuous at $z = b$, then $g(Z_n) \xrightarrow{p} g(b)$. ■

Similarly, it is almost as easy to show that if $Z_n \xrightarrow{p} b$ and $Y_n \xrightarrow{p} c$, and if $g(z, y)$ is continuous at $(z, y) = (b, c)$, then $g(Z_n, Y_n) \xrightarrow{p} g(b, c)$ (Exercise 16). Indeed, Theorem 6.2.5 extends to any finite number k of sequences that converge in probability and a continuous function of k variables.

The law of large numbers helps to explain why a histogram (Definition 3.7.9) can be used as an approximation to a p.d.f.

Theorem 6.2.6 Histograms. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Let $c_1 < c_2$ be two constants. Define $Y_i = 1$ if $c_1 \le X_i < c_2$ and $Y_i = 0$ if not. Then $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ is the proportion of $X_1, \ldots, X_n$ that lie in the interval $[c_1, c_2)$, and $\overline{Y}_n \xrightarrow{p} \Pr(c_1 \le X_1 < c_2)$.

Proof By construction, $Y_1, Y_2, \ldots$ are i.i.d. Bernoulli random variables with parameter $p = \Pr(c_1 \le X_1 < c_2)$. Theorem 6.2.4 says that $\overline{Y}_n \xrightarrow{p} p$. ■

In words, Theorem 6.2.6 says the following: If we draw a histogram with the area of the bar over each subinterval being the proportion of a random sample that lies in the corresponding subinterval, then the area of each bar converges in probability to the probability that a random variable from the sequence lies in the subinterval. If the sample is large, we would then expect the area of each bar to be close to the probability. The same idea applies to a conditionally i.i.d. (given $Z = z$) sample, with $\Pr(c_1 \le X_1 < c_2)$ replaced by $\Pr(c_1 \le X_1 < c_2 | Z = z)$.

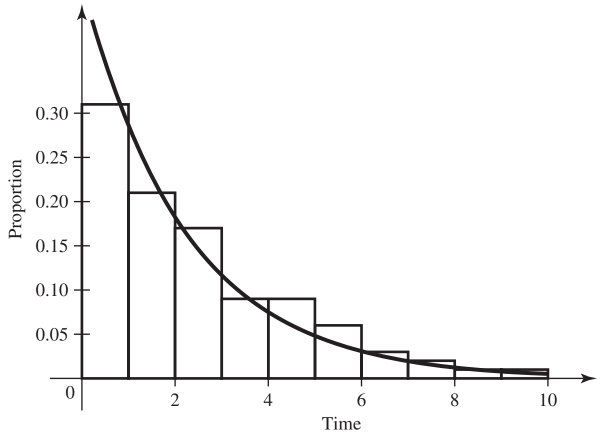


Figure 6.1 Histogram of service times for Example 6.2.4 together with graph of the conditional p.d.f. from which the service times were simulated.

Example 6.2.4

Rate of Service. In Example 3.7.20, we drew a histogram of an observed sample of $n = 100$ service times. The service times were actually simulated as an i.i.d. sample from the exponential distribution with parameter 0.446. Figure 6.1 reproduces the histogram overlayed with the graph of $g(x|z_0)$ where $z_0 = 0.446$. Because the width of each bar is 1, the area of each bar equals the proportion of the sample that lies in the corresponding interval. The area under the curve $g(x|z_0)$ is $\Pr(c_1 \le X_1 < c_2|Z = z_0)$ for each interval $[c_1, c_2)$. Notice how closely the area under the conditional p.d.f. matches the area of each bar. ◀

The reason that the p.d.f. and the heights of the bars in the histogram in Fig. 6.1 match so closely is that the area of each bar is converging in probablity to the area under the graph of the p.d.f. The sum of the areas of the bars is 1, which is the same as the area under the graph of the p.d.f. If we had chosen the heights of the bars in the histogram to represent counts, then the sum of the areas of the bars would have been $n = 100$, and the bars would have been about 100 times as high as the p.d.f.

We could choose a different width for the subintervals in the histogram and still keep the areas equal to the proportions in the subintervals.

Example 6.2.5

Rate of Service. In Example 6.2.4, we can choose 20 bars of width 0.5 instead of 10 bars of width 1. To make the area of each bar represent the proportion in the subinterval, the height of each bar should equal the proportion divided by 0.5. The probability of an observation being in each interval $[c_1, c_2)$ would be

$$\begin{aligned}\Pr(c_1 \le X_1 < c_2|Z = x) &= \int_{c_1}^{c_2} g(x|z)dx \approx (c_2 - c_1)g([c_1 + c_2]/2|z)\\ &= 0.5 * g([c_1 + c_2]/2|z). \qquad (6.2.6)\end{aligned}$$

Recall that the probability in (6.2.6) should be close to the proportion of the sample in the interval. If we divide both the probability and the proportion by 0.5, we see that the height of the histogram bar should be close to $g([c_1 + c_2]/2)$. Hence, the graph of the p.d.f. should still be close to the heights of the histogram bars. What we are doing here is choosing $r = n(b - a)/k$ in Defintion 3.7.9. Figure 6.2 shows the histogram with 20 intervals of length 0.5 together with the same p.d.f. from Fig. 6.1. The bar heights are still similar to the p.d.f., but they are much more variable in

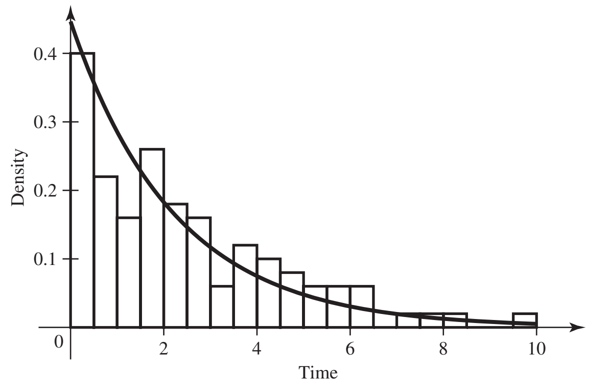


Figure 6.2 Modified histogram of service times from Example 6.2.4 together with graph of the conditional p.d.f. This time, the width of each interval is 0.5.

Fig. 6.2 compared to Fig. 6.1. Exercise 17 helps to explain why the bar heights are more variable in this example. ◀

The reasoning used to construct Figures 6.1 and 6.2 applies even when the subintervals used to construct the histogram have different widths. In this case, each bar should have height equal to the raw count divided by both n (the sample size) and the width of the corresponding subinterval.

Weak Laws and Strong Laws

There are other concepts of the convergence of a sequence of random variables, in addition to the concept of convergence in probability that has been presented above. For example, it is said that a sequence $Z_1, Z_2, \ldots$ *converges to a constant b with probability 1 if*

$$\Pr\left(\lim_{n\to\infty} Z_n = b\right) = 1.$$

A careful investigation of the concept of convergence with probability 1 is beyond the scope of this book. It can be shown that if a sequence $Z_1, Z_2, \ldots$ converges to b with probability 1, then the sequence will also converge to b in probability. For this reason, convergence with probability 1 is often called *strong convergence*, whereas convergence in probability is called *weak convergence*. In order to emphasize the distinction between these two concepts of convergence, the result that here has been called simply the law of large numbers is often called the *weak law of large numbers*. The *strong law of large numbers* can then be stated as follows: If $\overline{X}_n$ is the sample mean of a random sample of size n from a distribution with mean μ, then

$$\Pr\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

The proof of this result will not be given here. There are examples of sequences of random variables that converge in probability but that do not converge with probability 1. Exercise 22 is one such example. Another type of converges is *convergence in quadratic mean*, which is introduced in Exercises 10–13.

Chernoff Bounds

One way to think of the Chebyshev inequality is as an application of the Markov inequalitty to the random variable $(X-\mu)^2$. This idea generalizes to other functions and leads to a sharper bound on the probability in the tail of a distribution when the bound applies. Before giving the general result, we give a simple example to illustrate the potential improvement that it can provide.

Example 6.2.6 Binomial Random Variable. Suppose that X has the binomial distribution with parameters n and 1/2. We would like a bound to the probability that X/n is far from its mean 1/2. To be specific, suppose that we would like a bound for

$$\Pr\left(\left|\frac{X}{n}-\frac{1}{2}\right|\ge\frac{1}{10}\right). \tag{6.2.7}$$

The Chebyshev inequality gives the bound $\mathrm{Var}(X/n)/(1/10)^2$, which equals $25/n$.

Instead of applying the Chebyshev inequality, define $Y = X - n/2$ and rewrite the probability in (6.2.7) as the sum of the following two probabilities:

$$\Pr\left(\frac{X}{n}\ge\frac{1}{2}+\frac{1}{10}\right)=\Pr\left(Y\ge\frac{n}{10}\right),\quad\text{and}$$
$$\Pr\left(\frac{X}{n}\le\frac{1}{2}-\frac{1}{10}\right)=\Pr\left(-Y\ge\frac{n}{10}\right). \tag{6.2.8}$$

For each $s>0$, rewrite the first of the probabilities in (6.2.8) as

$$\Pr\left(Y\ge\frac{n}{10}\right)=\Pr\left[\exp(sY)\ge\exp\left(\frac{ns}{10}\right)\right]$$
$$\le\frac{E[\exp(sY)]}{\exp(ns/10)},$$

where the inequality follows from the Markov inequality. This equation involves the moment generating function of Y, $\psi(s)=E[\exp(sY)]$. The m.g.f. of Y can be found by applying Theorem 4.4.3 with $p=1/2$, $a=1$, and $b=-n/2$ together with Equation (5.2.4). The result is

$$\psi(s)=\left(\frac{1}{2}\left[\exp(s)+1\right]\exp(-s/2)\right)^n, \tag{6.2.9}$$

for all s. Let $s=1/2$ in (6.2.9) to obtain the bound

$$\Pr\left(Y\ge\frac{n}{10}\right)\le\psi(1/2)\exp(-n/20)$$
$$=\exp(-n/20)\left(\frac{1}{2}\left[\exp(1/2)+1\right]\exp(-1/4)\right)^n=0.9811^n.$$

Similarly, we can write the second probability in (6.2.8) as

$$\Pr\left(-Y\ge\frac{n}{10}\right)=\Pr\left[\exp(-sY)\ge\exp\left(\frac{ns}{10}\right)\right], \tag{6.2.10}$$

where $s>0$. The m.g.f. of $-Y$ is $\psi(-s)$. Let $s=1/2$ in (6.2.10) and apply the Markov inequality to obtatin the bound

$$\Pr\left(-Y \ge \frac{n}{10}\right) \le \psi(-1/2)\exp(-n/20)$$
$$= \exp(-n/20)\left(\frac{1}{2}\left[\exp(-1/2)+1\right]\exp(1/4)\right)^n = 0.9811^n.$$

Hence, we obtain the bound

$$\Pr\left(\left|\frac{X}{n} - \frac{1}{2}\right| \ge \frac{1}{10}\right) \le 2(0.9811)^n. \tag{6.2.11}$$

The bound in (6.2.11) decreases exponentially fast as n increases, while the Chebyshev bound $25/n$ decreases proportionally to $1/n$. For example, with $n = 100$, 200, 300, the Chebychev bounds are 0.25, 0.125, and 0.0833. The corresponding bounds from (6.2.11) are 0.2967, 0.0440, and 0.0065. ◀

The choice of $s = 1/2$ in Example 6.2.6 was arbitrary. Theorem 6.2.7 says that we can replace this arbitrary choice with the choice that leads to the smallest possible bound. The proof of Theorem 6.2.7 is a straightforward application of the Markov inequality. (See Exercise 18 in this section.)

Theorem 6.2.7 Chernoff Bounds. Let X be a random variable with moment generating function ψ. Then, for every real t,

$$\Pr(X \ge t) \le \min_{s>0} \exp(-st)\psi(s).$$ ■

Theorem 6.2.7 is most useful when X is the sum of n i.i.d. random variables each with finite m.g.f. and when $t = nu$ for a large value of n and some fixed u. This was the case in Example 6.2.6.

Example 6.2.7 Average of Geometric Random Sample. Suppose that $X_1, X_2, \ldots$ are i.i.d. geometric random variables with parameter p. We would like a bound to the probability that $\overline{X}_n$ is far from the mean $(1-p)/p$. To be specific, for each fixed $u > 0$, we would like a bound for

$$\Pr\left(\left|\overline{X}_n - \frac{1-p}{p}\right| \ge u\right). \tag{6.2.12}$$

Let $X = \sum_{i=1}^{n} X_i - n(1-p)/p$. For each $u > 0$, Theorem 6.2.7 can be used to bound both

$$\Pr\left(\overline{X}_n \ge \frac{1-p}{p} + u\right) = \Pr(X \ge nu), \quad \text{and}$$
$$\Pr\left(\overline{X}_n \le \frac{1-p}{p} - u\right) = \Pr(-X \ge nu).$$

Since (6.2.12) equals $\Pr(X \ge nu) + \Pr(-X \ge nu)$, the bound we seek is the sum of the two bounds that we get for $\Pr(X \ge nu)$ and $\Pr(-X \ge nu)$.

The m.g.f. of X can be found by applying Theorem 4.4.3 with $a = 1$ and $b = -n(1-p)/p$ together with Theorem 5.5.3. The result is

$$\psi(s) = \left(\frac{p\exp[-s(1-p)/p]}{1-(1-p)\exp(s)}\right)^n. \tag{6.2.13}$$

The m.g.f. of $-X$ is $\psi(-s)$. According to Theorem 6.2.7,

$$\Pr(X \ge nu) \le \min_{s>0} \psi(s)\exp(-snu). \tag{6.2.14}$$

We find the minimum of $\psi(s)\exp(-snu)$ by finding the minimum of its logarithm. Using (6.2.13), we get that

$$\log[\psi(s)\exp(-snu)] = n\left\{\log(p) - s\frac{1-p}{p} - \log[1-(1-p)\exp(s)] - su\right\}.$$

The deriviative of this expression with respect to s equals 0 at

$$s = -\log\left[\frac{(1+u)p+1-p}{up+1-p}(1-p)\right], \tag{6.2.15}$$

and the second derivative is positive. If $u > 0$, then the value of s in (6.2.15) is positive and $\psi(s)$ is finite. Hence, the value of s in (6.2.15) provides the minimum in (6.2.14). That minimum can be expressed as q^n where

$$q = [p(1+u)+1-p]\left[\frac{(1+u)p+1-p}{up+1-p}(1-p)\right]^{u+(1-p)/p} \tag{6.2.16}$$

and $0 < q < 1$. (See Exercise 19 for a proof.) Hence, $\Pr(X \ge nu) \le q^n$.

For $\Pr(-X \ge nu)$, we notice first that $\Pr(-X \ge nu) = 0$ if $u \ge (1-p)/p$ because $\sum_{i=1}^n X_i \ge 0$. If $u \ge (1-p)/p$, then the overall bound on (6.2.12) is q^n. For $0 < u < (1-p)/p$, the value of s that minimizes $\psi(-s)\exp(-snu)$ is

$$s = -\log\left[\frac{(1-u)p+1-p}{1-p-up}(1-p)\right],$$

which is positive when $0 < u < (1-p)/p$. The value of $\min_{s>0}\psi(-s)\exp(-snu)$ is r^n, where

$$r = [p(1-u)+1-p]\left[\frac{(1-u)p+1-p}{1-p-up}(1-p)\right]^{-u+(1-p)/p}$$

and $0 < r < 1$. Hence, the Chernoff bound is q^n if $u \ge (1-p)/p$ and is $q^n + r^n$ if $0 < u < (1-p)/p$. As such, the bound decreases exponenially fast as n increases. This is a marked impovement over the Chebyshev bound, which decreases like a constant over n. ◂

Summary

The law of large numbers says that the sample mean of a random sample converges in probability to the mean μ of the individual random variables, if the variance exists. This means that the sample mean will be close to μ if the size of the random sample is sufficiently large. The Chebyshev inequality provides a (crude) bound on how high the probability is that the sample mean will be close to μ. Chernoff bounds can be sharper, but are harder to compute.

Exercises

1. For each integer n, let X_n be a nonnegative random variable with finite mean μ_n. Prove that if $\lim_{n\to\infty}\mu_n = 0$, then $X_n \xrightarrow{p} 0$.

2. Suppose that X is a random variable for which

$$\Pr(X \ge 0) = 1 \text{ and } \Pr(X \ge 10) = 1/5.$$

Prove that $E(X) \ge 2$.

3. Suppose that X is a random variable for which $E(X) = 10$, $\Pr(X \le 7) = 0.2$, and $\Pr(X \ge 13) = 0.3$. Prove that $\mathrm{Var}(X) \ge 9/2$.

4. Let X be a random variable for which $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Construct a probability distribution for X such that

$$\Pr(|X-\mu| \ge 3\sigma) = 1/9.$$

5. How large a random sample must be taken from a given distribution in order for the probability to be at least 0.99 that the sample mean will be within 2 standard deviations of the mean of the distribution?

6. Suppose that $X_1, \dots, X_n$ form a random sample of size n from a distribution for which the mean is 6.5 and the variance is 4. Determine how large the value of n must be in order for the following relation to be satisfied:

$$\Pr(6 \le \overline{X}_n \le 7) \ge 0.8.$$

7. Suppose that X is a random variable for which $E(X) = \mu$ and $E[(X-\mu)^4] = \beta_4$. Prove that

$$\Pr(|X-\mu| \ge t) \le \frac{\beta_4}{t^4}.$$

8. Suppose that 30 percent of the items in a large manufactured lot are of poor quality. Suppose also that a random sample of n items is to be taken from the lot, and let Q_n denote the proportion of the items in the sample that are of poor quality. Find a value of n such that $\Pr(0.2 \le Q_n \le 0.4) \ge 0.75$ by using **(a)** the Chebyshev inequality and **(b)** the tables of the binomial distribution at the end of this book.

9. Let $Z_1, Z_2, \dots$ be a sequence of random variables, and suppose that, for $n = 1, 2, \dots$, the distribution of Z_n is as follows:

$$\Pr(Z_n = n^2) = \frac{1}{n} \quad \text{and} \quad \Pr(Z_n = 0) = 1 - \frac{1}{n}.$$

Show that

$$\lim_{n\to\infty} E(Z_n) = \infty \text{ but } Z_n \xrightarrow{p} 0.$$

10. It is said that a sequence of random variables $Z_1, Z_2, \dots$ *converges to a constant b in quadratic mean if*

$$\lim_{n\to\infty} E[(Z_n - b)^2] = 0. \tag{6.2.17}$$

Show that Eq. (6.2.17) is satisfied if and only if

$$\lim_{n\to\infty} E(Z_n) = b \quad \text{and} \quad \lim_{n\to\infty} \text{Var}(Z_n) = 0.$$

Hint: Use Exercise 5 of Sec. 4.3.

11. Prove that if a sequence $Z_1, Z_2, \dots$ converges to a constant b in quadratic mean, then the sequence also converges to b in probability.

12. Let $\overline{X}_n$ be the sample mean of a random sample of size n from a distribution for which the mean is μ and the variance is σ^2, where $\sigma^2 < \infty$. Show that $\overline{X}_n$ converges to μ in quadratic mean as $n \to \infty$.

13. Let $Z_1, Z_2, \dots$ be a sequence of random variables, and suppose that for $n = 2, 3, \dots$, the distribution of Z_n is as follows:

$$\Pr\left(Z_n = \frac{1}{n}\right) = 1 - \frac{1}{n^2} \quad \text{and} \quad \Pr(Z_n = n) = \frac{1}{n^2}.$$

a. Does there exist a constant c to which the sequence converges in probability?

b. Does there exist a constant c to which the sequence converges in quadratic mean?

14. Let f be a p.f. for a discrete distribution. Suppose that $f(x) = 0$ for $x \notin [0, 1]$. Prove that the variance of this distribution is at most 1/4. *Hint:* Prove that there is a distribution supported on just the two points {0, 1} that has variance at least as large as f does and then prove that the variance of a distribution supported on {0, 1} is at most 1/4.

15. Prove Theorem 6.2.5.

16. Suppose that $Z_n \xrightarrow{p} b$, $Y_n \xrightarrow{p} c$, and $g(z, y)$ is a function that is continuous at $(z, y) = (b, c)$. Prove that $g(Z_n, Y_n)$ converges in probability to $g(b, c)$.

17. Let X have the binomial distribution with parameters n and p. Let Y have the binomial distribution with parameters n and p/k with $k > 1$. Let $Z = kY$.

a. Show that X and Z have the same mean.

b. Find the variances of X and Z. Show that, if p is small, then the variance of Z is approximately k times as large as the variance of X.

c. Show why the results above explain the higher variability in the bar heights in Fig. 6.2 compared to Fig. 6.1.

18. Prove Theorem 6.2.7.

19. Return to Example 6.2.7.

a. Prove that the $\min_{s>0} \psi(s) \exp(-snu)$ equals q^n, where q is given in (6.2.16).

b. Prove that $0 < q < 1$. *Hint:* First, show that $0 < q < 1$ if $u = 0$. Next, let $x = up + 1 - p$ and show that $\log(q)$ is a decreasing function of x.

20. Return to Example 6.2.6. Find the Chernoff bound for the probability in (6.2.7).

21. Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables having the exponential distribution with parameter 1. Let $Y_n = \sum_{i=1}^n X_i$ for each $n = 1, 2, \dots$.

a. For each $u > 1$, compute the Chernoff bound on $\Pr(Y_n > nu)$.

b. What goes wrong if we try to compute the Chernoff bound when $u < 1$?

22. In this exercise, we construct an example of a sequence of random variables Z_n such that $Z_n \xrightarrow{p} 0$ but

$$\Pr\left(\lim_{n\to\infty} Z_n = 0\right) = 0. \tag{6.2.18}$$

That is, Z_n converges in probability to 0, but Z_n does not converge to 0 with probability 1. Indeed, Z_n converges to 0 with probability 0.

Let X be a random variable having the uniform distribution on the interval [0, 1]. We will construct a sequence of functions $h_n(x)$ for $n = 1, 2, \ldots$ and define $Z_n = h_n(X)$. Each function h_n will take only two values, 0 and 1. The set of x where $h_n(x) = 1$ is determined by dividing the interval [0, 1] into k nonoverlappling subintervals of length $1/k$ for $k = 1, 2, \ldots$, arranging these intervals in sequence, and letting $h_n(x) = 1$ on the nth interval in the sequence for $n = 1, 2, \ldots$. For each k, there are k nonoverlapping subintervals, so the number of subintervals with lengths $1, 1/2, 1/3, \ldots, 1/k$ is

$$1 + 2 + 3 + \cdots + k = \frac{k(k+1)}{2}.$$

The remainder of the construction is based on this formula. The first interval in the sequence has length 1, the next two have length 1/2, the next three have length 1/3, etc.

a. For each $n = 1, 2, \ldots$, prove that there is a unique positive integer k_n such that

$$\frac{(k_n - 1)k_n}{2} < n \le \frac{k_n(k_n + 1)}{2}.$$

b. For each $n = 1, 2, \ldots$, let $j_n = n - (k_n - 1)k_n/2$. Show that j_n takes the values $1, \ldots, k_n$ as n runs through $1 + (k_n - 1)k_n/2, \ldots, k_n(k_n + 1)/2$.

c. Define

$$h_n(x) = \begin{cases} 1 & \text{if } (j_n - 1)/k_n \le x < j_n/k_n, \\ 0 & \text{if not.} \end{cases}$$

Show that, for every $x \in [0, 1)$, $h_n(x) = 1$ for one and only one n among $1 + (k_n - 1)k_n/2, \ldots, k_n(k_n + 1)/2$.

d. Show that $Z_n = h_n(X)$ takes the value 1 infinitely often with probability 1.

e. Show that (6.2.18) holds.

f. Show that $\Pr(Z_n = 0) = 1 - 1/k_n$ and $\lim_{n\to\infty} k_n = \infty$.

g. Show that $Z_n \xrightarrow{p} 0$.

23. Prove that the sequence of random variables Z_n in Exercise 22 converges in quadratic mean (definition in Exercise 10) to 0.

24. In this exercise, we construct an example of a sequence of random variables Z_n such that Z_n converges to 0 with probability 1, but Z_n fails to converge to 0 in quadratic mean. Let X be a random variable having the uniform distribution on the interval [0, 1]. Define the sequence Z_n by $Z_n = n^2$ if $0 < X < 1/n$ and $Z_n = 0$ otherwise.

a. Prove that Z_n converges to 0 with probability 1.

b. Prove that Z_n does not converge to 0 in quadratic mean.

6.3 The Central Limit Theorem

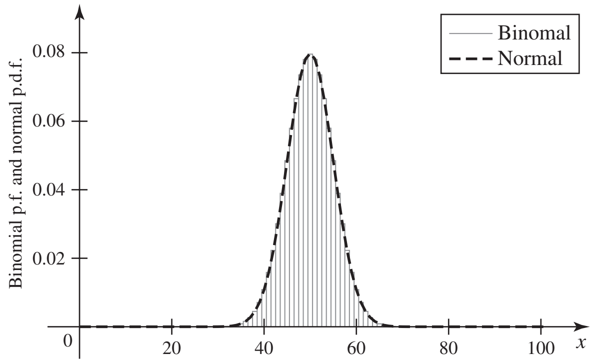
The sample mean of a large random sample of random variables with mean μ and finite variance σ^2 has approximately the normal distribution with mean μ and variance σ^2/n. This result helps to justify the use of the normal distribution as a model for many random variables that can be thought of as being made up of many independent parts. Another version of the central limit theorem is given that applies to independent random variables that are not identically distributed. We also introduce the delta method, which allows us to compute approximate distributions for functions of random variables.

Statement of the Theorem

Example 6.3.1

A Large Sample. A clinical trial has 100 patients who will receive a treatment. Patients who don't receive the treatment survive for 18 months with probability 0.5 each. We assume that all patients are independent. The trial is to see whether the new treatment can increase the probability of survival significantly. Let X be the number of patients out of the 100 who survive for 18 months. If the probabiity of success were 0.5 for the patients on the treatment (the same as without the treatment), then X would have the binomial distribution with parameters $n = 100$ and $p = 0.5$. The p.f. of X is graphed as a bar chart with the solid line in Fig. 6.3. The shape of the bar chart is reminiscent of a bell-shaped curve. The normal p.d.f. with the same mean $\mu = 50$ and variance $\sigma^2 = 25$ as the binomial distribution is also graphed with the dotted line. ◀

Figure 6.3 Comparison of the binomial p.f. with parameters 100 and 0.5 to the normal p.d.f. with mean 50 and variance 25.



In Examples 5.4.1 and 5.4.2, we illustrated how the Poisson distribution provides a good approximation to a binomial distribution with a large n and small p. Example 6.3.1 shows how a normal distribution can be a good approximation to a binomial distribution with a large n and not so small p. The central limit theorem (Theorem 6.3.1) is a formal statement of how normal distributions can approximate distributions of general sums or averages of i.i.d. random variables.

In Corollary 5.6.2, we saw that if a random sample of size n is taken from the normal distribution with mean μ and variance σ^2, then the sample average $\overline{X}_n$ has the normal distribution with mean μ and variance σ^2/n. The simple version of the central limit theorem that we give in this section says that whenever a random sample of size n is taken from *any* distribution with mean μ and variance σ^2, the sample average $\overline{X}_n$ will have a distribution that is *approximately* normal with mean μ and variance σ^2/n.

This result was established for a random sample from a Bernoulli distribution by A. de Moivre in the early part of the eighteenth century. The proof for a random sample from an arbitrary distribution was given independently by J. W. Lindeberg and P. Lévy in the early 1920s. A precise statement of their theorem will be given now, and an outline of the proof of that theorem will be given later in this section. We shall also state another central limit theorem pertaining to the sum of independent random variables that are not necessarily identically distributed and shall present some examples illustrating both theorems.

Theorem 6.3.1 Central Limit Theorem (Lindeberg and Lévy). If the random variables $X_1, \dots, X_n$ form a random sample of size n from a given distribution with mean μ and variance σ^2 ($0 < \sigma^2 < \infty$), then for each fixed number x,

$$\lim_{n\to\infty} \Pr\left[\frac{\overline{X}_n - \mu}{\sigma/n^{1/2}} \le x\right] = \Phi(x), \tag{6.3.1}$$

where Φ denotes the c.d.f. of the standard normal distribution. ■

The interpretation of Eq. (6.3.1) is as follows: If a large random sample is taken from any distribution with mean μ and variance σ^2, regardless of whether this distribution is discrete or continuous, then the distribution of the random variable $n^{1/2}(\overline{X}_n - \mu)/\sigma$ will be approximately the standard normal distribution. Therefore, the distribution of $\overline{X}_n$ will be approximately the normal distribution with mean μ and variance σ^2/n, or, equivalently, the distribution of the sum $\sum_{i=1}^{n} X_i$ will be

approximately the normal distribution with mean $n\mu$ and variance $n\sigma^2$. It is in this last form that the central limit theorem was illustrated in Example 6.3.1.

Example 6.3.2

Tossing a Coin. Suppose that a fair coin is tossed 900 times. We shall approximate the probability of obtaining more than 495 heads.

For $i = 1, \ldots, 900$, let $X_i = 1$ if a head is obtained on the ith toss and let $X_i = 0$ otherwise. Then $E(X_i) = 1/2$ and $\text{Var}(X_i) = 1/4$. Therefore, the values $X_1, \ldots, X_{900}$ form a random sample of size $n = 900$ from a distribution with mean 1/2 and variance 1/4. It follows from the central limit theorem that the distribution of the total number of heads $H = \sum_{i=1}^{900} X_i$ will be approximately the normal distribution for which the mean is $(900)(1/2) = 450$, the variance is $(900)(1/4) = 225$, and the standard deviation is $(225)^{1/2} = 15$. Therefore, the variable $Z = (H - 450)/15$ will have approximately the standard normal distribution. Thus,

$$\begin{aligned}\Pr(H > 495) &= \Pr\left(\frac{H - 450}{15} > \frac{495 - 450}{15}\right)\\ &= \Pr(Z > 3) \approx 1 - \Phi(3) = 0.0013.\end{aligned}$$ ◀

The exact probability 0.0012 to four decimal places.

Example 6.3.3

Sampling from a Uniform Distribution. Suppose that a random sample of size $n = 12$ is taken from the uniform distribution on the interval [0, 1]. We shall approximate the value of $\Pr\left(\left|\overline{X}_n - \frac{1}{2}\right| \le 0.1\right)$.

The mean of the uniform distribution on the interval [0, 1] is 1/2, and the variance is 1/12 (see Exercise 3 of Sec. 4.3). Since $n = 12$ in this example, it follows from the central limit theorem that the distribution of $\overline{X}_n$ will be approximately the normal distribution with mean 1/2 and variance 1/144. Therefore, the distribution of the variable $Z = 12\left(\overline{X}_n - \frac{1}{2}\right)$ will be approximately the standard normal distribution. Hence,

$$\begin{aligned}\Pr\left(\left|\overline{X}_n - \frac{1}{2}\right| \le 0.1\right) &= \Pr\left[12\left|\overline{X}_n - \frac{1}{2}\right| \le 1.2\right]\\ &= \Pr(|Z| \le 1.2) \approx 2\Phi(1.2) - 1 = 0.7698.\end{aligned}$$

For the special case of $n = 12$, the random variable Z has the form $Z = \sum_{i=1}^{12} X_i - 6$. At one time, some computers produced standard normal pseudo-random numbers by adding 12 uniform pseudo-random numbers and subtracting 6. ◀

Example 6.3.4

Poisson Random Variables. Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with mean θ. Let $\overline{X}_n$ be the average. Then $\mu = \theta$ and $\sigma^2 = \theta$. The central limit theorem says that $n^{1/2}(\overline{X}_n - \theta)/\theta^{1/2}$ has approximately the standard normal distribution. In particular, the central limit theorem says that $\overline{X}_n$ should be close to μ with high probability. The probability that $|\overline{X}_n - \theta|$ is less than some small number c could be approximated using the standard normal c.d.f.:

$$\Pr\left(|\overline{X}_n - \theta| < c\right) \approx 2\Phi\left(cn^{1/2}\theta^{-1/2}\right) - 1. \qquad (6.3.2)$$

◀

The type of convergence that appears in the central limit theorem, specifically, Eq. (6.3.1), arises in other contexts and has a special name.

Definition 6.3.1 Convergence in Distribution/Asymptotic Distribution. Let $X_1, X_2, \ldots$ be a sequence of random variables, and for $n = 1, 2, \ldots$, let F_n denote the c.d.f. of X_n. Also, let F^* be a c.d.f. Then it is said that the sequence $X_1, X_2, \ldots$ *converges in distribution* to F^* if

$$\lim_{n\to\infty} F_n(x) = F^*(x), \tag{6.3.3}$$

for all x at which $F^*(x)$ is continuous. Sometimes, it is simply said that X_n converges in distribution to F^*, and F^* is called the *asymptotic distribution* of X_n. If F^* has a name, then we say that X_n converges in distribution to that name.

Thus, according to Theorem 6.3.1, as indicated in Eq. (6.3.1), the random variable $n^{1/2}(\overline{X}_n - \mu)/\sigma$ converges in distribution to the standard normal distribution, or, equivalently, the asymptotic distribution of $n^{1/2}(\overline{X}_n - \mu)/\sigma$ is the standard normal distribution.

Effect of the Central Limit Theorem The central limit theorem provides a plausible explanation for the fact that the distributions of many random variables studied in physical experiments are approximately normal. For example, a person's height is influenced by many random factors. If the height of each person is determined by adding the values of these individual factors, then the distribution of the heights of a large number of persons will be approximately normal. In general, the central limit theorem indicates that the distribution of the sum of many random variables can be approximately normal, even though the distribution of each random variable in the sum differs from the normal.

Example 6.3.5 Determining a Simulation Size. In Example 6.2.2 on page 351, an environmental engineer wanted to determine the size of a simulation to estimate the mean proportion of water contaminant that was lead. Use of the Chebyshev inequality in that example suggested that a simulation of size 2,000,000 will guarantee that the estimate will be less than 0.005 away from the true mean proportion with probability at least 0.98. In this example, we shall use the central limit theorem to determine a much smaller simulation size that should still provide the same accuracy bound. The estimate of the mean proportion will be the average $\overline{R}_n$ of all of the simulated proportions $R_1, \ldots, R_n$ from the n simulations that will be run. As we noted in Example 6.2.2, the variance of each R_i is $\sigma^2 \le 1$, and hence the central limit theorem says that $\overline{R}_n$ has approximately the normal distribution with mean equal to the true mean proportion $E(R_i)$ and variance at most $1/n$. Since the probability of being close to the mean decreases as the variance increases, we see that

$$\begin{aligned}
\Pr(|\overline{R}_n - E(R_i)| < 0.005) &\approx \Phi\left(\frac{0.005}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{-0.005}{\sigma/\sqrt{n}}\right) \\
&\ge \Phi\left(\frac{0.005}{1/\sqrt{n}}\right) - \Phi\left(\frac{-0.005}{1/\sqrt{n}}\right) \\
&= 2\Phi(0.005\sqrt{n}) - 1.
\end{aligned}$$

If we set $2\Phi(0.005\sqrt{n}) - 1 = 0.98$, we obtain

$$n = \frac{1}{0.005^2}\Phi^{-1}(0.99)^2 = 40{,}000 \times 2.326^2 = 216{,}411.$$

That is, we only need a little more than 10 percent of the simulation size that the Chebyshev inequality suggested. (Since σ^2 is actually no more than 1/4, we really only need n =54,103. See Exercise 14 in Sec. 6.2 for a proof that a discrete distribution on

the interval [0, 1] can have variance at most 1/4. The continuous case is slightly more complicated, but also true.) ◀

Other Examples of Convergence in Distribution In Chapter 5, we saw three examples of limit theorems involving discrete distributions. Theorems 5.3.4, 5.4.5, and 5.4.6 all showed that a sequence of p.f.'s converged to some other p.f. In Exercise 7 in Sec. 6.5, you can prove a general result that implies that the three theorems just mentioned are examples of convergence in distribution.

The Delta Method

Example 6.3.6 Rate of Service. Customers arrive at a queue for service, and the ith customer is served in some time X_i after reaching the head of the queue. If we assume that $X_1, \dots, X_n$ form a random sample of service times with mean μ and finite variance σ^2, we might be interested in using $1/\overline{X}_n$ to estimate the rate of service. The central limit theorem tells us something about the approximate distribution of $\overline{X}_n$ if n is large, but what can we say about the distribution of $1/\overline{X}_n$? ◀

Suppose that $X_1, \dots, X_n$ form a random sample from a distribution that has finite mean μ and finite variance σ^2. The central limit theorem says that $n^{1/2}(\overline{X}_n - \mu)/\sigma$ has approximately the standard normal distribution. Now suppose that we are interested in the distribution of some function α of $\overline{X}_n$. We shall assume that α is a differentiable function whose derivative is nonzero at μ. We shall approximate the distribution of $\alpha(\overline{X}_n)$ by a method known in statistics as the *delta method*.

Theorem 6.3.2 Delta Method. Let $Y_1, Y_2, \dots$ be a sequence of random variables, and let F^* be a continuous c.d.f. Let θ be a real number, and let $a_1, a_2, \dots$ be a sequence of positive numbers that increase to ∞. Suppose that $a_n(Y_n - \theta)$ converges in distribution to F^*. Let α be a function with continuous derivative such that $\alpha'(\theta) \neq 0$. Then $a_n[\alpha(Y_n) - \alpha(\theta)]/\alpha'(\theta)$ converges in distribution to F^*.

Proof We shall give only an outline of the proof. Because $a_n \to \infty$, Y_n must get close to θ with high probability as $n \to \infty$. If not, $|a_n(Y_n - \theta)|$ would go to ∞ with nonzero probability and then the c.d.f. of $a_n(Y_n - \theta)$ would not converge to a c.d.f. Because α is continuous, $\alpha(Y_n)$ must also be close to $\alpha(\theta)$ with high probability. Therefore, we shall use a Taylor series expansion of $\alpha(Y_n)$ around θ,

$$\alpha(Y_n) \approx \alpha(\theta) + \alpha'(\theta)(Y_n - \theta), \tag{6.3.4}$$

where we have ignored all terms involving $(Y_n - \theta)^2$ and higher powers. Subtract $\alpha(\theta)$ from both sides of Eq. (6.3.4), and then multiply both sides by $a_n/\alpha'(\theta)$ to get

$$\frac{a_n}{\alpha'(\theta)}(Y_n - \theta) \approx a_n(Y_n - \theta). \tag{6.3.5}$$

We then conclude that the distribution of the left side of Eq. (6.3.5) will be approximately the same as the distribution of the right side of the equation, which is approximately F^*. ■

The most common application of Theorem 6.3.2 occurs when Y_n is the average of a random sample from a distribution with finite variance. We state that case in the following corollary.

Corollary 6.3.1 Delta Method for Average of a Random Sample. Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables from a distribution with mean μ and finite variance σ^2. Let α

be a function with continuous derivative such that $\alpha'(\mu) \neq 0$. Then the asymptotic distribution of

$$\frac{n^{1/2}}{\sigma\alpha'(\mu)}[\alpha(\overline{X}_n) - \alpha(\mu)]$$

is the standard normal distribution.

Proof Apply Theorem 6.3.2 with $Y_n = \overline{X}_n$, $a_n = n^{1/2}/\sigma$, $\theta = \mu$, and F^* being the standard normal c.d.f. ■

A common way to report the result in Corollary 6.3.1 is to say that the distribution of $\alpha(\overline{X}_n)$ is approximately the normal distribution with mean $\alpha(\mu)$ and variance $\sigma^2[\alpha'(\mu)]^2/n$.

Example 6.3.7

Rate of Service. In Example 6.3.6, we are interested in the distribution of $\alpha(\overline{X}_n)$ where $\alpha(x) = 1/x$ for $x > 0$. We can apply the delta method by finding $\alpha'(x) = -1/x^2$. It follows that the asymptotic distribution of

$$-\frac{n^{1/2}\mu^2}{\sigma}\left(\frac{1}{\overline{X}_n} - \frac{1}{\mu}\right)$$

is the standard normal distribution. Alternatively, we might say that $1/\overline{X}_n$ has approximately the normal distribution with mean $1/\mu$ and variance $\sigma^2/[n\mu^4]$. ◄

Variance Stabilizing Transformations If we were to observe a random sample of Poisson random variables as in Example 6.3.4, we would assume that θ is unknown. In such a case we cannot compute the probability in Eq. (6.3.2), because the approximate variance of $\overline{X}_n$ depends on θ. For this reason, it is sometimes desirable to transform $\overline{X}_n$ by a function α so that the approximate distribution of $\alpha(\overline{X}_n)$ has a variance that is a known value. Such a function is called a *variance stabilizing transformation*. We can often find a variance stabilizing transformation by running the delta method in reverse. In general, we note that the approximate distribution of $\alpha(\overline{X}_n)$ has variance $\alpha'(\mu)^2\sigma^2/n$. In order to make this variance constant, we need $\alpha'(\mu)$ to be a constant times $1/\sigma$. If σ^2 is a function $g(\mu)$, then we achieve this goal by letting

$$\alpha(\mu) = \int_a^\mu \frac{dx}{g(x)^{1/2}}, \tag{6.3.6}$$

where a is an arbitrary constant that makes the integral finite.

Example 6.3.8

Poisson Random Variables. In Example 6.3.4, we have $\sigma^2 = \theta = \mu$, so that $g(\mu) = \mu$. According to Eq. (6.3.6), we should let

$$\alpha(\mu) = \int_0^\mu \frac{dx}{x^{1/2}} = 2\mu^{1/2}.$$

It follows that $2\overline{X}_n^{1/2}$ has approximately the normal distribution with mean $2\theta^{1/2}$ and variance $1/n$. For each number $c > 0$, we have

$$\Pr\left(|2\overline{X}_n^{1/2} - 2\theta^{1/2}| < c\right) \approx 2\Phi\left(cn^{1/2}\right) - 1. \tag{6.3.7}$$

In Chapter 8, we shall see how to use Eq (6.3.7) to estimate θ when we assume that θ is unknown. ◄

The Central Limit Theorem (Liapounov) for the Sum of Independent Random Variables

We shall now state a central limit theorem that applies to a sequence of random variables $X_1, X_2, \dots$ that are independent but not necessarily identically distributed. This theorem was first proved by A. Liapounov in 1901. We shall assume that $E(X_i) = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2$ for $i = 1, \dots, n$. Also, we shall let

$$Y_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}}. \tag{6.3.8}$$

Then $E(Y_n) = 0$ and $\text{Var}(Y_n) = 1$. The theorem that is stated next gives a sufficient condition for the distribution of this random variable Y_n to be approximately the standard normal distribution.

Theorem 6.3.3 Suppose that the random variables $X_1, X_2, \dots$ are independent and that $E(|X_i - \mu_i|^3) < \infty$ for $i = 1, 2, \dots$ Also, suppose that

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n E\left(|X_i - \mu_i|^3\right)}{\left(\sum_{i=1}^n \sigma_i^2\right)^{3/2}} = 0. \tag{6.3.9}$$

Finally, let the random variable Y_n be as defined in Eq. (6.3.8). Then, for each fixed number x,

$$\lim_{n\to\infty} \Pr(Y_n \le x) = \Phi(x). \tag{6.3.10}$$

■

The interpretation of this theorem is as follows: If Eq. (6.3.9) is satisfied, then for every large value of n, the distribution of $\sum_{i=1}^n X_i$ will be approximately the normal distribution with mean $\sum_{i=1}^n \mu_i$ and variance $\sum_{i=1}^n \sigma_i^2$. It should be noted that when the random variables $X_1, X_2, \dots$ are identically distributed and the third moments of the variables exist, Eq. (6.3.9) will automatically be satisfied and Eq. (6.3.10) then reduces to Eq. (6.3.1).

The distinction between the theorem of Lindeberg and Lévy and the theorem of Liapounov should be emphasized. The theorem of Lindeberg and Lévy applies to a sequence of i.i.d. random variables. In order for this theorem to be applicable, it is sufficient to assume only that the variance of each random variable is finite. The theorem of Liapounov applies to a sequence of independent random variables that are not necessarily identically distributed. In order for this theorem to be applicable, it must be assumed that the third moment of each random variable is finite and satisfies Eq. (6.3.9).

The Central Limit Theorem for Bernoulli Random Variables By applying the theorem of Liapounov, we can establish the following result.

Theorem 6.3.4 Suppose that the random variables $X_1, \dots, X_n$ are independent and X_i has the Bernoulli distribution with parameter p_i ($i = 1, 2, \dots$). Suppose also that the infinite series $\sum_{i=1}^\infty p_i(1 - p_i)$ is divergent, and let

$$Y_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n p_i}{\left(\sum_{i=1}^n p_i(1 - p_i)\right)^{1/2}}. \tag{6.3.11}$$

Then for every fixed number x,

$$\lim_{n\to\infty} \Pr(Y_n \le x) = \Phi(x). \tag{6.3.12}$$

Proof Here $\Pr(X_i = 1) = p_i$ and $\Pr(X_i = 0) = 1 - p_i$. Therefore,

$$E(X_i) = p_i, \ \mathrm{Var}(X_i) = p_i(1-p_i),$$

$$\begin{aligned} E\left(|X_i - p_i|^3\right) &= p_i(1-p_i)^3 + (1-p_i)p_i^3 = p_i(1-p_i)\left(p_i^2 + (1-p_i^2)\right) \\ &\le p_i(1-p_i), \end{aligned} \tag{6.3.13}$$

It follows that

$$\frac{\sum_{i=1}^n E\left(|X_i - p_i|^3\right)}{\left(\sum_{i=1}^n p_i(1-p_i)\right)^{3/2}} \le \frac{1}{\left(\sum_{i=1}^n p_i(1-p_i)\right)^{1/2}}. \tag{6.3.14}$$

Since the infinite series $\sum_{i=1}^\infty p_i(1-p_i)$ is divergent, then $\sum_{i=1}^n p_i(1-p_i) \to \infty$ as $n \to \infty$, and it can be seen from the relation (6.3.14) that Eq. (6.3.9) will be satisfied. In turn, it follows from Theorem 6.3.3 that Eq. (6.3.10) will be satisfied. Since Eq. (6.3.12) is simply a restatement of Eq. (6.3.10) for the particular random variables being considered here, the proof of the theorem is complete. ■

Theorem 6.3.4 implies that if the infinite series $\sum_{i=1}^\infty p_i(1-p_i)$ is divergent, then the distribution of the sum $\sum_{i=1}^n X_i$ of a large number of independent Bernoulli random variables will be approximately the normal distribution with mean $\sum_{i=1}^n p_i$ and variance $\sum_{i=1}^n p_i(1-p_i)$. It should be kept in mind, however, that a typical practical problem will involve only a finite number of random variables $X_1, \dots, X_n$, rather than an infinite sequence of random variables. In such a problem, it is not meaningful to consider whether or not the infinite series $\sum_{i=1}^\infty p_i(1-p_i)$ is divergent, because only a finite number of values $p_1, \dots, p_n$ will be specified in the problem. In a certain sense, therefore, the distribution of the sum $\sum_{i=1}^n X_i$ can *always* be approximated by a normal distribution. The critical question is whether or not this normal distribution provides a *good* approximation to the actual distribution of $\sum_{i=1}^n X_i$. The answer depends, of course, on the values of $p_1, \dots, p_n$.

Since the normal distribution will be attained more and more closely as $\sum_{i=1}^n p_i(1-p_i) \to \infty$, the normal distribution provides a good approximation when the value of $\sum_{i=1}^n p_i(1-p_i)$ is large. Furthermore, since the value of each term $p_i(1-p_i)$ is a maximum when $p_i = 1/2$, the approximation will be best when n is large and the values of $p_1, \dots, p_n$ are close to 1/2.

Example 6.3.9 Examination Questions. Suppose that an examination contains 99 questions arranged in a sequence from the easiest to the most difficult. Suppose that the probability that a particular student will answer the first question correctly is 0.99, the probability that he will answer the second question correctly is 0.98, and, in general, the probability that he will answer the ith question correctly is $1 - i/100$ for $i = 1, \dots, 99$. It is assumed that all questions will be answered independently and that the student must answer at least 60 questions correctly to pass the examination. We shall determine the probability that the student will pass.

Let $X_i = 1$ if the ith question is answered correctly and $X_i = 0$ otherwise. Then $E(X_i) = p_i = 1 - (i/100)$ and $\mathrm{Var}(X_i) = p_i(1-p_i) = (i/100)[1 - (i/100)]$. Also,

$$\sum_{i=1}^{99} p_i = 99 - \frac{1}{100}\sum_{i=1}^{99} i = 99 - \frac{1}{100}\cdot\frac{(99)(100)}{2} = 49.5$$

and

$$\sum_{i=1}^{99} p_i(1-p_i) = \frac{1}{100}\sum_{i=1}^{99} i - \frac{1}{(100)^2}\sum_{i=1}^{99} i^2$$
$$= 49.5 - \frac{1}{(100)^2}\cdot\frac{(99)(100)(199)}{6} = 16.665.$$

It follows from the central limit theorem that the distribution of the total number of questions that are answered correctly, which is $\sum_{i=1}^{99} X_i$, will be approximately the normal distribution with mean 49.5 and standard deviation $(16.665)^{1/2} = 4.08$. Therefore, the distribution of the variable

$$Z = \frac{\sum_{i=1}^{n} X_i - 49.5}{4.08}$$

will be approximately the standard normal distribution. It follows that

$$\Pr\left(\sum_{i=1}^{n} X_i \geq 60\right) = \Pr(Z \geq 2.5735) \simeq 1 - \Phi(2.5735) = 0.0050. \quad ◄$$

Outline of Proof of Central Limit Theorem

Convergence of the Moment Generating Functions Moment generating functions are important in the study of convergence in distribution because of the following theorem, the proof of which is too advanced to be presented here.

Theorem 6.3.5 Let $X_1, X_2, \ldots$ be a sequence of random variables. For $n = 1, 2, \ldots$, let F_n denote the c.d.f. of X_n, and let ψ_n denote the m.g.f. of X_n.

Also, let X^* denote another random variable with c.d.f. F^* and m.g.f. ψ^*. Suppose that the m.g.f.'s ψ_n and ψ^* exist ($n = 1, 2, \ldots$). If $\lim_{n\to\infty} \psi_n(t) = \psi^*(t)$ for all values of t in some interval around the point $t = 0$, then the sequence $X_1, X_2, \ldots$ converges in distribution to X^*. ■

In other words, the sequence of c.d.f.'s $F_1, F_2, \ldots$ must converge to the c.d.f. F^* if the corresponding sequence of m.g.f.'s $\psi_1, \psi_2, \ldots$ converges to the m.g.f. ψ^*.

Outline of the Proof of Theorem 5.7.1 We are now ready to outline a proof of Theorem 6.3.1, which is the central limit theorem of Lindeberg and Lévy. We shall assume that the variables $X_1, \ldots, X_n$ form a random sample of size n from a distribution with mean μ and variance σ^2. We shall also assume, for convenience, that the m.g.f. of this distribution exists, although the central limit theorem is true even without this assumption.

For $i = 1, \ldots, n$, let $Y_i = (X_i - \mu)/\sigma$. Then the random variables $Y_1, \ldots, Y_n$ are i.i.d., and each has mean 0 and variance 1. Furthermore, let

$$Z_n = \frac{n^{1/2}(\overline{X}_n - \mu)}{\sigma} = \frac{1}{n^{1/2}}\sum_{i=1}^{n} Y_i.$$

We shall show that Z_n converges in distribution to a random variable having the standard normal distribution, as indicated in Eq. (6.3.1), by showing that the m.g.f. of Z_n converges to the m.g.f. of the standard normal distribution.

If $\psi(t)$ denotes the m.g.f. of each random variable Y_i $(i = 1, \ldots, n)$, then it follows from Theorem 4.4.4 that the m.g.f. of the sum $\sum_{i=1}^{n} Y_i$ will be $[\psi(t)]^n$. Also, it follows from Theorem 4.4.3 that the m.g.f. $\zeta_n(t)$ of Z_n will be

$$\zeta_n(t) = \left[\psi\left(\frac{t}{n^{1/2}}\right)\right]^n.$$

In this problem, $\psi'(0) = E(Y_i) = 0$ and $\psi''(0) = E(Y_i^2) = 1$. Therefore, the Taylor series expansion of $\psi(t)$ about the point $t = 0$ has the following form:

$$\begin{aligned}\psi(t) &= \psi(0) + t\psi'(0) + \frac{t^2}{2!}\psi''(0) + \frac{t^3}{3!}\psi'''(0) + \cdots \\ &= 1 + \frac{t^2}{2} + \frac{t^3}{3!}\psi'''(0) + \cdots.\end{aligned}$$

Also,

$$\zeta_n(t) = \left[1 + \frac{t^2}{2n} + \frac{t^3\psi'''(0)}{3!n^{3/2}} + \cdots\right]^n. \tag{6.3.15}$$

Apply Theorem 5.3.3 with $1 + a_n/n$ equal to the expression inside brackets in (6.3.15) and $c_n = n$. Since

$$\lim_{n\to\infty}\left[\frac{t^2}{2} + \frac{t^3\psi'''(0)}{3!n^{1/2}} + \cdots\right] = \frac{t^2}{2}.$$

it follows that

$$\lim_{n\to\infty} \zeta_n(t) = \exp\left(\frac{1}{2}t^2\right). \tag{6.3.16}$$

Since the right side of Eq. (6.3.16) is the m.g.f. of the standard normal distribution, it follows from Theorem 6.3.5 that the asymptotic distribution of Z_n must be the standard normal distribution.

An outline of the proof of the central limit theorem of Liapounov can also be given by proceeding along similar lines, but we shall not consider this problem further here.

Summary

Two versions of the central limit theorem were given. They conclude that the distribution of the average of a large number of independent random variables is close to a normal distribution. One theorem requires that the random variables all have the same distribution with finite variance. The other theorem does not require that the random variables be identically distributed, but instead requires that their third moments exist and satisfy condition (6.3.9). The delta method lets us find the approximate distribution of a smooth function of a sample average.

Exercises

1. Each minute a machine produces a length of rope with mean of 4 feet and standard deviation of 5 inches. Assuming that the amounts produced in different minutes are independent and identically distributed, approximate the probability that the machine will produce at least 250 feet in one hour.

2. Suppose that 75 percent of the people in a certain metropolitan area live in the city and 25 percent of the people live in the suburbs. If 1200 people attending a certain concert represent a random sample from the metropolitan area, what is the probability that the number of people from the suburbs attending the concert will be fewer than 270?

3. Suppose that the distribution of the number of defects on any given bolt of cloth is the Poisson distribution with mean 5, and the number of defects on each bolt is counted for a random sample of 125 bolts. Determine the probability that the average number of defects per bolt in the sample will be less than 5.5.

4. Suppose that a random sample of size n is to be taken from a distribution for which the mean is μ and the standard deviation is 3. Use the central limit theorem to determine approximately the smallest value of n for which the following relation will be satisfied:

$$\Pr(|\overline{X}_n - \mu| < 0.3) \geq 0.95.$$

5. Suppose that the proportion of defective items in a large manufactured lot is 0.1. What is the smallest random sample of items that must be taken from the lot in order for the probability to be at least 0.99 that the proportion of defective items in the sample will be less than 0.13?

6. Suppose that three girls A, B, and C throw snowballs at a target. Suppose also that girl A throws 10 times, and the probability that she will hit the target on any given throw is 0.3; girl B throws 15 times, and the probability that she will hit the target on any given throw is 0.2; and girl C throws 20 times, and the probability that she will hit the target on any given throw is 0.1. Determine the probability that the target will be hit at least 12 times.

7. Suppose that 16 digits are chosen at random with replacement from the set $\{0, \ldots, 9\}$. What is the probability that their average will lie between 4 and 6?

8. Suppose that people attending a party pour drinks from a bottle containing 63 ounces of a certain liquid. Suppose also that the expected size of each drink is 2 ounces, that the standard deviation of each drink is 1/2 ounce, and that all drinks are poured independently. Determine the probability that the bottle will not be empty after 36 drinks have been poured.

9. A physicist makes 25 independent measurements of the specific gravity of a certain body. He knows that the limitations of his equipment are such that the standard deviation of each measurement is σ units.

a. By using the Chebyshev inequality, find a lower bound for the probability that the average of his measurements will differ from the actual specific gravity of the body by less than $\sigma/4$ units.

b. By using the central limit theorem, find an approximate value for the probability in part (a).

10. A random sample of n items is to be taken from a distribution with mean μ and standard deviation σ.

a. Use the Chebyshev inequality to determine the smallest number of items n that must be taken in order to satisfy the following relation:

$$\Pr\left(|\overline{X}_n - \mu| \leq \frac{\sigma}{4}\right) \geq 0.99.$$

b. Use the central limit theorem to determine the smallest number of items n that must be taken in order to satisfy the relation in part (a) approximately.

11. Suppose that, on the average, 1/3 of the graduating seniors at a certain college have two parents attend the graduation ceremony, another third of these seniors have one parent attend the ceremony, and the remaining third of these seniors have no parents attend. If there are 600 graduating seniors in a particular class, what is the probability that not more than 650 parents will attend the graduation ceremony?

12. Let X_n be a random variable having the binomial distribution with parameters n and p_n. Assume that $\lim_{n\to\infty} np_n = \lambda$. Prove that the m.g.f. of X_n converges to the m.g.f. of the Poisson distribution with mean λ.

13. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution with unknown mean θ and variance σ^2. Assuming that $\theta \neq 0$, determine the asymptotic distribution of $\overline{X}_n^3$.

14. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution with mean 0 and unknown variance σ^2.

a. Determine the asymptotic distribution of the statistic $\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)^{-1}$.

b. Find a variance stabilizing transformation for the statistic $\frac{1}{n}\sum_{i=1}^n X_i^2$.

15. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables each having the uniform distribution on the interval $[0, \theta]$ for some real number $\theta > 0$. For each n, define Y_n to be the maximum of $X_1, \ldots, X_n$.

a. Show that the c.d.f. of Y_n is

$$F_n(y) = \begin{cases} 0 & \text{if } x \le 0, \\ (y/\theta)^n & \text{if } 0 < y < \theta, \\ 1 & \text{if } y > \theta. \end{cases}$$

Hint: Read Example 3.9.6.

b. Show that $Z_n = n(Y_n - \theta)$ converges in distribution to the distribution with c.d.f.

$$F^*(z) = \begin{cases} \exp(z/\theta) & \text{if } z < 0, \\ 1 & \text{if } z > 0. \end{cases}$$

Hint: Apply Theorem 5.3.3 after finding the c.d.f. of Z_n.

c. Use Theorem 6.3.2 to find the approximate distribution of Y_n^2 when n is large.

6.4 The Correction for Continuity

Some applications of the central limit theorem allow us to approximate the probability that a discrete random variable X lies in an interval $[a, b]$ by the probability that a normal random variable lies in that interval. The approximation can be improved slightly by being careful about how we approximate $\Pr(X = a)$ *and* $\Pr(X = b)$.

Approximating a Discrete Distribution by a Continuous Distribution

Example 6.4.1 **A Large Sample.** In Example 6.3.1, we illustrated how the normal distribution with mean 50 and variance 25 could approximate the distribution of a random variable X that has the binomial distribution with parameters 100 and 0.5. In particular, if Y has the normal distribution with mean 50 and variance 25, we know that $\Pr(Y \le x)$ is close to $\Pr(X \le x)$ for all x. But the approximation has some systematic errors. Figure 6.4 shows the two c.d.f.'s over the range $30 \le x < 70$. The two c.d.f.'s are very close at $x = n + 0.5$ for each integer n. But for each integer n, $\Pr(Y \le x) < \Pr(X \le x)$ for x a little above n and $\Pr(Y \le x) > \Pr(X \le x)$ for x a little below n. We ought to be able to make use of these systematic discrepancies in order to improve the approximation. ◄

Suppose that X has a discrete distribution that can be approximated by a normal distribution, such as in Example 6.4.1. In this section, we shall describe a standard method for improving the quality of such an approximation based on the systematic discrepancies that were noted at the end of Example 6.4.1.

Let $f(x)$ be the p.f. of the discrete random variable X, and suppose that we wish to approximate the distribution of X by a continuous distribution with p.d.f. $g(x)$. To

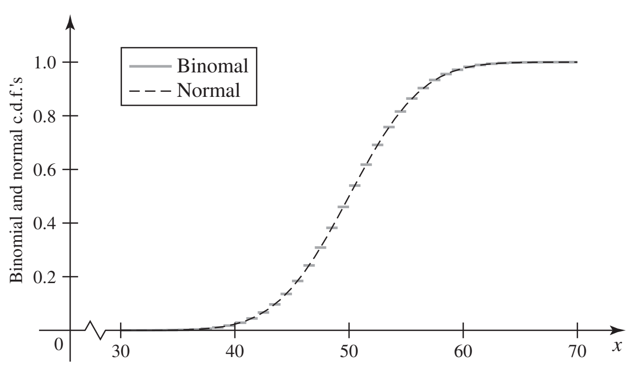


Figure 6.4 Comparison of binomial and normal c.d.f.'s.

aid the discussion, let Y be a random variable with p.d.f. g. Also, for simplicity, we shall assume that all of the possible values of X are integers. This condition is satisfied for the binomial, hypergeometric, Poisson, and negative binomial distributions described in this text.

If the distribution of Y provides a good approximation to the distribution of X, then for all integers a and b, we can approximate the discrete probability

$$\Pr(a \le X \le b) = \sum_{x=a}^{b} f(x) \tag{6.4.1}$$

by the continuous probability

$$\Pr(a \le Y \le b) = \int_a^b g(x)\, dx. \tag{6.4.2}$$

Indeed, this approximation was used in Examples 6.3.2 and 6.3.9, where $g(x)$ was the appropriate normal p.d.f. derived from the central limit theorem.

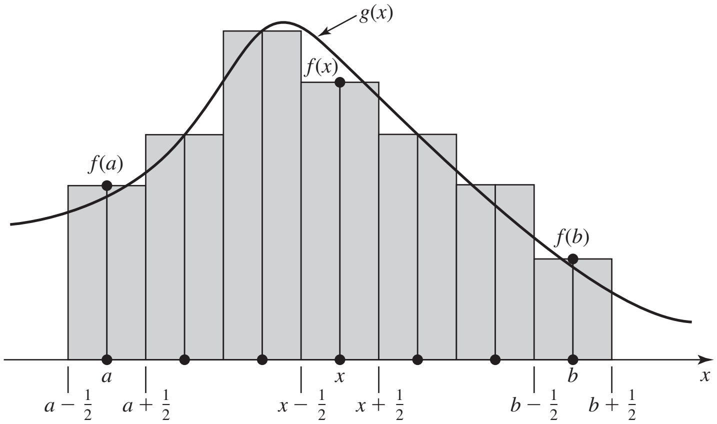
This simple approximation has the following shortcoming: Although $\Pr(X \ge a)$ and $\Pr(X > a)$ will typically have different values for the discrete distribution of X, $\Pr(Y \ge a) = \Pr(Y > a)$ because Y has a continuous distribution. Another way of expressing this shortcoming is as follows: Although $\Pr(X = x) > 0$ for each integer x that is a possible value of X, $\Pr(Y = x) = 0$ for all x.

Approximating a Bar Chart

The p.f. $f(x)$ of a discrete random variable X can be represented by a *bar chart*, as sketched in Fig. 6.5. For each integer x, the probability of $\{X = x\}$ is represented by the area of a rectangle with a base that extends from $x - \frac{1}{2}$ to $x + \frac{1}{2}$ and with a height $f(x)$. Thus, the area of the rectangle for which the center of the base is at the integer x is simply $f(x)$. An approximating p.d.f. $g(x)$ is also sketched in Fig. 6.5. A bar chart with areas of bars proportional to probabilities is analogous to a histogram (see page 165) with areas of bars proportional to proportions of a sample.

From this point of view, it can be seen that $\Pr(a \le X \le b)$, as specified in Eq. (6.4.1), is the sum of the areas of the rectangles in Fig. 6.5 that are centered at $a, a + 1, \ldots, b$. It can also be seen from Fig. 6.5 that the sum of these areas is

Figure 6.5 Approximating a bar chart by using a p.d.f.



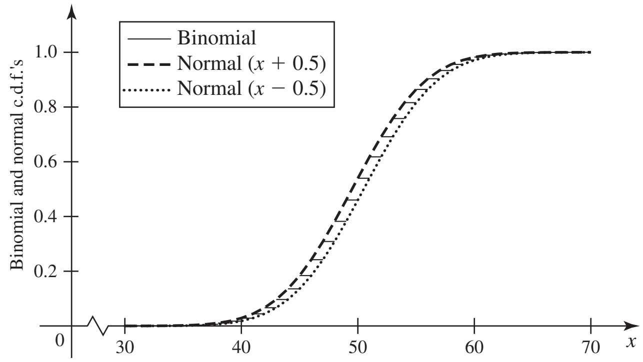


Figure 6.6 Comparison of binomial c.d.f. with normal c.d.f. shifted to the right and to the left by 0.5.

approximated by the integral

$$\Pr(a - 1/2 < Y < b + 1/2) = \int_{a-(1/2)}^{b+(1/2)} g(x)\,dx. \tag{6.4.3}$$

The adjustment from the integral in (6.4.2) to the integral in (6.4.3) is called the *correction for continuity*.

Example 6.4.2

A Large Sample. At the end of Example 6.4.1, we found that when x was a little above an integer, the approximating probability $\Pr(Y \le x)$ is a bit smaller than the actual probability $\Pr(X \le x)$. The correction for continuity shifts the c.d.f. of Y to the left by 0.5 when we want to compute $\Pr(Y \le x)$ for x a little above an integer. This shift replaces $\Pr(Y \le x)$ by $\Pr(Y \le x + 0.5)$, which is larger and usually closer to $\Pr(X \le x)$. Similarly, when we want to compute $\Pr(Y \le x)$ when x is a little below an integer, the correction for continuity shifts the c.d.f. of Y to the right by 0.5 which replaces $\Pr(Y \le x)$ by $\Pr(Y \le x - 0.5)$. Figure 6.6 illustrates both of these shifts and shows how they each approximate the actual binomial c.d.f. better than the unshifted normal c.d.f. in Fig. 6.4. ◄

If we use the correction for continuity, we find that the probability $f(a)$ of the single integer a can be approximated as follows:

$$\begin{aligned}\Pr(X = a) &= \Pr\left(a - \frac{1}{2} \le X \le a + \frac{1}{2}\right)\\ &\approx \int_{a-(1/2)}^{a+(1/2)} g(x)\,dx.\end{aligned} \tag{6.4.4}$$

Similarly,

$$\begin{aligned}\Pr(X > a) = \Pr(X \ge a + 1) &= \Pr\left(X \ge a + \frac{1}{2}\right)\\ &\approx \int_{a+(1/2)}^{\infty} g(x)\,dx.\end{aligned} \tag{6.4.5}$$

Example 6.4.3

Examination Questions. To illustrate the use of the correction for continuity, we shall again consider Example 6.3.9. In that example, an examination contains 99 questions of varying difficulty and it is desired to determine $\Pr(X \ge 60)$, where X denotes the total number of questions that a particular student answers correctly. Then, under the conditions of the example, it is found from the central limit theorem that the discrete

distribution of X could be approximated by the normal distribution with mean 49.5 and standard deviation 4.08. Let $Z = (X - 49.5)/4.08$.

If we use the correction for continuity, we obtain

$$\Pr(X \geq 60) = \Pr(X \geq 59.5) = \Pr\left(Z \geq \frac{59.5 - 49.5}{4.08}\right)$$
$$\approx 1 - \Phi(2.4510) = 0.007.$$

This value is somewhat larger than the value 0.005, which was obtained in Sec. 6.3, without the correction. ◀

Example 6.4.4

Coin Tossing. Suppose that a fair coin is tossed 20 times and that all tosses are independent. What is the probability of obtaining exactly 10 heads?

Let X denote the total number of heads obtained in the 20 tosses. According to the central limit theorem, the distribution of X will be approximately the normal distribution with mean 10 and standard deviation $[(20)(1/2)(1/2)]^{1/2} = 2.236$. If we use the correction for continuity,

$$\Pr(X = 10) = \Pr(9.5 \leq X \leq 10.5)$$
$$= \Pr\left(-\frac{0.5}{2.236} \leq Z \leq \frac{0.5}{2.236}\right)$$
$$\approx \Phi(0.2236) - \Phi(-0.2236) = 0.177.$$

The exact value of $\Pr(X = 10)$ found from the table of binomial probabilities given at the back of this book is 0.1762. Thus, the normal approximation with the correction for continuity is quite good. ◀

Summary

Let X be a random variable that takes only integer values. Suppose that X has approximately the normal distribution with mean μ and variance σ^2. Let a and b be integers, and suppose that we wish to approximate $\Pr(a \leq X \leq b)$. The correction to the normal distribution approximation for continuity is to use $\Phi([b + 1/2 - \mu]/\sigma) - \Phi([a - 1/2 - \mu]/\sigma)$ rather than $\Phi([b - \mu]/\sigma) - \Phi([a - \mu]/\sigma)$ as the approximation.

Exercises

1. Let $X_1, \dots, X_{30}$ be independent random variables each having a discrete distribution with p.f.

$$f(x) = \begin{cases} 1/4 & \text{if } x = 0 \text{ or } 2, \\ 1/2 & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Use the central limit theorem and the correction for continuity to approximate the probability that $X_1 + \cdots + X_{30}$ is at most 33.

2. Let X denote the total number of successes in 15 Bernoulli trials, with probability of success $p = 0.3$ on each trial.

a. Determine approximately the value of $\Pr(X = 4)$ by using the central limit theorem with the correction for continuity.

b. Compare the answer obtained in part (a) with the exact value of this probability.

3. Using the correction for continuity, determine the probability required in Example 6.3.2.

4. Using the correction for continuity, determine the probability required in Exercise 2 of Sec. 6.3.

5. Using the correction for continuity, determine the probability required in Exercise 3 of Sec. 6.3.

6. Using the correction for continuity, determine the probability required in Exercise 6 of Sec. 6.3.

7. Using the correction for continuity, determine the probability required in Exercise 7 of Sec. 6.3.

6.5 Supplementary Exercises

1. Suppose that a pair of balanced dice are rolled 120 times, and let X denote the number of rolls on which the sum of the two numbers is 7. Use the central limit theorem to determine a value of k such that $\Pr(|X - 20| \le k)$ is approximately 0.95.

2. Suppose that X has a Poisson distribution with a very large mean λ. Explain why the distribution of X can be approximated by the normal distribution with mean λ and variance λ. In other words, explain why $(X - \lambda)/\lambda^{1/2}$ converges in distribution, as $\lambda \to \infty$, to a random variable having the standard normal distribution.

3. Suppose that X has the Poisson distribution with mean 10. Use the central limit theorem, both without and with the correction for continuity, to determine an approximate value for $\Pr(8 \le X \le 12)$. Use the table of Poisson probabilities given in the back of this book to assess the quality of these approximations.

4. Suppose that X is a random variable such that $E(X^k)$ exists and $\Pr(X \ge 0) = 1$. Prove that for $k > 0$ and $t > 0$,

$$\Pr(X \ge t) \le \frac{E(X^k)}{t^k}.$$

5. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with parameter p. Let $\overline{X}_n$ be the sample average. Find a variance stabilizing transformation for $\overline{X}_n$. *Hint:* When trying to find the integral of $(p[1 - p])^{-1/2}$, make the substitution $z = \sqrt{p}$ and then think about arcsin, the inverse of the sin function.

6. Suppose that $X_1, \ldots, X_n$ form a random sample from the exponential distribution with mean θ. Let $\overline{X}_n$ be the sample average. Find a variance stabilizing transformation for $\overline{X}_n$.

7. Suppose that $X_1, X_2, \ldots$ is a sequence of positive integer-valued random variables. Suppose that there is a function f such that for every $m = 1, 2, \ldots$, $\lim_{n\to\infty} \Pr(X_n = m) = f(m)$, $\sum_{m=1}^{\infty} f(m) = 1$, and $f(x) = 0$ for every x that is not a positive integer. Let F be the discrete c.d.f. whose p.f. is f. Prove that X_n converges in distribution to F.

8. Let $\{p_n\}_{n=1}^{\infty}$ be a sequence of numbers such that $0 < p_n < 1$ for all n. Assume that $\lim_{n\to\infty} p_n = p$ with $0 < p < 1$. Let X_n have the binomial distribution with parameters k and p_n for some positive integer k. Prove that X_n converges in distribution to the binomial distribution with parameters k and p.

9. Suppose that the number of minutes required to serve a customer at the checkout counter of a supermarket has an exponential distribution for which the mean is 3. Using the central limit theorem, approximate the probability that the total time required to serve a random sample of 16 customers will exceed one hour.

10. Suppose that we model the ocurrence of defects on a fabric manufacturing line as a Poisson process with rate 0.01 per square foot. Use the central limit theorem (both with and without the correction for continuity) to approximate the probability that one would find at least 15 defects in 2000 square feet of fabric.

11. Let X have the gamma distribution with parameters n and 3, where n is a large integer.

a. Explain why one can use the central limit theorem to approximate the distribution of X by a normal distribution.

b. Which normal distribution approximates the distribution of X?

12. Let X have the negative binomial distribution with parameters n and 0.2, where n is a large integer.

a. Explain why one can use the central limit theorem to approximate the distribution of X by a normal distribution.

b. Which normal distribution approximates the distribution of X?

Chapter 7

ESTIMATION

7.1 Statistical Inference

Recall our various clinical trial examples. What would we say is the probability that a future patient will respond successfully to treatment after we observe the results from a collection of other patients? This is the kind of question that statistical inference is designed to address. In general, statistical inference consists of making probabilistic statements about unknown quantities. For example, we can compute means, variances, quantiles, probabilities, and some other quantities yet to be introduced concerning unobserved random variables and unknown parameters of distributions. Our goal will be to say what we have learned about the unknown quantities after observing some data that we believe contain relevant information. Here are some other examples of questions that statistical inference can try to answer. What can we say about whether a machine is functioning properly after we observe some of its output? In a civil lawsuit, what can we say about whether there was discrimination after observing how different ethnic groups were treated? The methods of statistical inference, which we shall develop to address these questions, are built upon the theory of probability covered in the earlier chapters of this text.

Probability and Statistical Models

In the earlier chapters of this book, we discussed the theory and methods of probability. As new concepts in probability were introduced, we also introduced examples of the use of these concepts in problems that we shall now recognize as *statistical inference*. Before discussing statistical inference formally, it is useful to remind ourselves of those probability concepts that will underlie inference.

Example 7.1.1 Lifetimes of Electronic Components. A company sells electronic components and they are interested in knowing as much as they can about how long each component is likely to last. They can collect data on components that have been used under typical conditions. They choose to use the family of exponential distributions to model the length of time (in years) from when a component is put into service until it fails. They would like to model the components as all having the same failure rate θ, but there is uncertainty about the specific numerical value of θ. To be more precise,

let $X_1, X_2, \ldots$ stand for a sequence of component lifetimes in years. The company believes that if they knew the failure rate θ, then $X_1, X_2, \ldots$ would be i.i.d. random variables having the exponential distribution with parameter θ. (See Sec. 5.7 for the definition of exponential distributions. We are using the symbol θ for the parameter of our exponential distributions rather than β to match the rest of the notation in this chapter.) Suppose that the data that the company will observe consist of the values of $X_1, \ldots, X_m$ but that they are still interested in $X_{m+1}, X_{m+2}, \ldots$. They are also interested in θ because it is related to the average lifetime. As we saw in Eq. (5.7.17), the mean of an exponential random variable with parameter θ is $1/\theta$, which is why the company thinks of θ as the failure rate.

We imagine an experiment whose outcomes are sequences of lifetimes as described above. As mentioned already, if we knew the value θ, then $X_1, X_2, \ldots$ would be i.i.d. random variables. In this case, the law of large numbers (Theorem 6.2.4) says that the average $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to the mean $1/\theta$. And Theorem 6.2.5 says that $n/\sum_{i=1}^{n} X_i$ converges in probability to θ. Because θ is a function of the sequence of lifetimes that constitute each experimental outcome, it can be treated as a random variable. Suppose that, before observing the data, the company believes that the failure rate is probably around 0.5/year but there is quite a bit of uncertainty about it. They model θ as a random variable having the gamma distribution with parameters 1 and 2. To rephrase what was stated earlier, they also model $X_1, X_2, \ldots$ as conditionally i.i.d. exponential random variables with parameter θ given θ. They hope to learn more about θ from examining the sample data $X_1, \ldots, X_m$. They can never learn θ precisely, because that would require observing the entire infinite sequence $X_1, X_2, \ldots$. For this reason, θ is only hypothetically observable. ◀

Example 7.1.1 illustrates several features that will be common to most statistical inference problems and which constitute what we call a statistical model.

Definition 7.1.1 Statistical Model. A *statistical model* consists of an identification of random variables of interest (both observable and only hypothetically observable), a specification of a joint distribution or a family of possible joint distributions for the observable random variables, the identification of any parameters of those distributions that are assumed unknown and possibly hypothetically observable, and (if desired) a specification for a (joint) distribution for the unknown parameter(s). When we treat the unknown parameter(s) θ as random, then the joint distribution of the observable random variables indexed by θ is understood as the conditional distribution of the observable random variables given θ.

In Example 7.1.1, the observable random variables of interest form the sequence $X_1, X_2, \ldots$, while the failure rate θ is hypothetically observable. The family of possible joint distributions of $X_1, X_2, \ldots$ is indexed by the parameter θ. The joint distribution of the observables corresponding to the value θ is that $X_1, X_2, \ldots$ are i.i.d. random variables each having the exponential distribution with parameter θ. This is also the conditional distribution of $X_1, X_2, \ldots$ given θ because we are treating θ as a radom variable. The distribution of θ is the gamma distribution with parameters 1 and 2.

Note: Redefining Old Ideas. The reader will notice that a statistical model is nothing more than a formal identification of many features that we have been using in various examples throughout the earlier chapters of this book. Some examples need only a few of the features that make up a complete specification of a statistical model, while other examples use the complete specification. In Sections 7.1–7.4, we shall

introduce a considerable amount of terminology, most of which is mere formalization of concepts that have been introduced and used in several places earlier in the book. The purpose of all of this formalism is to help us to keep the concepts organized so that we can tell when we are applying the same ideas in new ways and when we are introducing new ideas.

We are now ready formally to introduce statistical inference.

Definition 7.1.2 Statistical Inference. A *statistical inference* is a procedure that produces a probabilistic statement about some or all parts of a statistical model.

By a "probabilistic statement" we mean a statement that makes use of any of the concepts of probability theory that were discussed earlier in the text or are yet to be discussed later in the text. Some examples include a mean, a conditional mean, a quantile, a variance, a conditional distribution for a random variable given another, the probability of an event, a conditional probability of an event given something, and so on. In Example 7.1.1, here are some examples of statistical inferences that one might wish to make:

- Produce a random variable Y (a function of $X_1, \ldots, X_m$) such that $\Pr(Y \geq \theta|\theta) = 0.9$.
- Produce a random variable Y that we expect to be close to θ.
- Compute how likely it is that the average of the next 10 lifetimes, $\frac{1}{10}\sum_{i=m+1}^{m+10} X_i$, is at least 2.
- Say something about how confident we are that $\theta \leq 0.4$ after observing $X_1, \ldots, X_m$.

All of these types of inference and others will be discussed in more detail later in this book.

In Definition 7.1.1, we distinguished between observable and hypothetically observable random variables. We reserved the name *observable* for a random variable that we are essentially certain that we could observe if we devoted the necessary effort to observe it. The name *hypothetically observable* was used for a random variable that would require infinite resources to observe, such as the limit (as $n \to \infty$) of the sample averages of the first n observables. In this text, such hypothetically observable random variables will correspond to the parameters of the joint distribution of the observables as in Example 7.1.1. Because these parameters figure so prominently in many of the types of inference problems that we will see, it pays to formalize the concept of parameter.

Definition 7.1.3 Parameter/Parameter space. In a problem of statistical inference, a characteristic or combination of characteristics that determine the joint distribution for the random variables of interest is called a *parameter* of the distribution. The set Ω of all possible values of a parameter θ or of a vector of parameters $(\theta_1, \ldots, \theta_k)$ is called the *parameter space*.

All of the families of distributions introduced earlier (and to be introduced later) in this book have parameters that are included in the names of the individual members of the family. For example, the family of binomial distributions has parameters that we called n and p, the family of normal distributions is parameterized by the mean μ and variance σ^2 of each distribution, the family of uniform distributions on intervals is parameterized by the endpoints of the intervals, the family of exponential distributions is parameterized by the rate parameter θ, and so on.

In Example 7.1.1, the parameter θ (the failure rate) must be positive. Therefore, unless certain positive values of θ can be explicitly ruled out as possible values of θ, the parameter space Ω will be the set of all positive numbers. As another example, suppose that the distribution of the heights of the individuals in a certain population is assumed to be the normal distribution with mean μ and variance σ^2, but that the exact values of μ and σ^2 are unknown. The mean μ and the variance σ^2 determine the particular normal distribution for the heights of individuals. So (μ, σ^2) can be considered a pair of parameters. In this example of heights, both μ and σ^2 must be positive. Therefore, the parameter space Ω can be taken as the set of all pairs (μ, σ^2) such that $\mu > 0$ and $\sigma^2 > 0$. If the normal distribution in this example represents the distribution of the heights in inches of the individuals in some particular population, we might be certain that $30 < \mu < 100$ and $\sigma^2 < 50$. In this case, the parameter space Ω could be taken as the smaller set of all pairs (μ, σ^2) such that $30 < \mu < 100$ and $0 < \sigma^2 < 50$.

The important feature of the parameter space Ω is that it must contain all possible values of the parameters in a given problem, in order that we can be certain that the actual value of the vector of parameters is a point in Ω.

Example 7.1.2

A Clinical Trial. Suppose that 40 patients are going to be given a treatment for a condition and that we will observe for each patient whether or not they recover from the condition. We are most likely also intersted in a large collection of additional patients besides the 40 to be observed. To be specific, for each patient $i = 1, 2, \ldots,$ let $X_i = 1$ if patient i recovers, and let $X_i = 0$ if not. As a collection of possible distributions for $X_1, X_2, \ldots,$ we could choose to say that the X_i are i.i.d. having the Bernoulli distribution with parameter p for $0 \le p \le 1$. In this case, the parameter p is known to lie in the closed interval [0, 1], and this interval could be taken as the parameter space. Notice also that the law of large numbers (Theorem 6.2.4) says that p is the limit as n goes to infinity of the proportion of the first n patients who recover. ◀

In most problems, there is a natural interpretation for the parameter as a feature of the possible distributions of our data. In Example 7.1.2, the parameter p has a natural interpretation as the proportion out of a large population of patients given the treatment who recover from the condition. In Example 7.1.1, the parameter θ has a natural interpretation as a failure rate, that is, one over the average lifetime of a large population of lifetimes. In such cases, inference about parameters can be interpreted as inference about the feature that the parameter represents. In this text, all parameters will have such natural interpretations. In examples that one encounters outside of an introductory course, interpretations may not be as straightforward.

Examples of Statistical Inference

Here are some of the examples of statistical models and inferences that were introduced earlier in the text.

Example 7.1.3

A Clinical Trial. The clinical trial introduced in Example 2.1.4 was concerned with how likely patients are to avoid relapse while under various treatments. For each i, let $X_i = 1$ if patient i in the imipramine group avoids relapse and $X_i = 0$ otherwise. Let P stand for the proportion of patients who avoid relapse out of a large group receiving imipramine treatment. If P is unknown, we can model $X_1, X_2, \ldots$ as i.i.d.

Bernoulli random variables with parameter p conditional on $P = p$. The patients in the imipramine column of Table 2.1 should provide us with some information that changes our uncertainty about P. A statistical inference would consist of making a probability statement about the data and/or P, and what the data and P tell us about each other. For instance, in Example 4.7.8, we assumed that P had the uniform distribution on the interval [0, 1], and we found the conditional distribution of P given the observed results of the study. We also computed the conditional mean of P given the study results as well as the M.S.E. for trying to predict P both before and after observing the results of the study. ◀

Example 7.1.4 Radioactive Particles. In Example 5.7.8, radioactive particles reach a target according to a Poisson process with unknown rate β. In Exercise 22 of Sec. 5.7, you were asked to find the conditional distribution of β after observing the Poisson process for a certain amount of time. ◀

Example 7.1.5 Anthropometry of Flea Beetles. In Example 5.10.2, we plotted two physical measurements from a sample of 31 flea beetles together with contours of a bivariate normal distribution. The family of bivariate normal distributions is parameterized by five quantities: the two means, the two variances, and the correlation. The choice of which set of five parameters to use for the fitted distribution is a form of statistical inference known as *estimation*. ◀

Example 7.1.6 Interval for Mean. Suppose that the heights of men in a certain population follow the normal distribution with mean μ and variance 9, as in Example 5.6.7. This time, assume that we do not know the value of the mean μ, but rather we wish to learn about it by sampling from the population. Suppose that we decide to sample $n = 36$ men and let $\overline{X}_n$ stand for the average of their heights. Then the interval $\left(\overline{X}_n - 0.98, \overline{X}_n + 0.98\right)$ computed in Example 5.6.8 has the property that it will contain the value of μ with probability 0.95. ◀

Example 7.1.7 Discrimination in Jury Selection. In Example 5.8.4, we were interested in whether there was evidence of discrimination against Mexican Americans in juror selection. Figure 5.8 shows how people who came into the case with different opinions about the extent of discrimination (if any) could alter their opinions in the light of learning the numerical evidence presented in the case. ◀

Example 7.1.8 Service Times in a Queue. Suppose that customers in a queue must wait for service, and that we get to observe the service times of several customers. Suppose that we are interested in the rate at which customers are served. In Example 5.7.3, we let Z stand for the service rate, and in Example 5.7.4, we showed how to find the conditional distribution of Z given several observed service times. ◀

General Classes of Inference Problems

Prediction One form of inference is to try to predict random variables that have not yet been observed. In Example 7.1.1, we might be interested in the average of the next 10 lifetimes, $\frac{1}{10}\sum_{i=m+1}^{m+10} X_i$. In the clinical trial example (Example 7.1.3), we might be interested in predicting how many patients from the next set of patients in the imipramine group will have successful outcome. In virtually every statistical inference problem, in which we have not observed all of the relevant data, prediction

is possible. When the unobserved quantity to be predicted is a parameter, prediction is usually called *estimation,* as in Example 7.1.5.

Statistical Decision Problems In many statistical inference problems, after the experimental data have been analyzed, we must choose a decision from some available class of decisions with the property that the consequences of each available decision depend on the unknown value of some parameter. For example, we might have to estimate the unknown failure rate θ of our electronic components when the consequences depend on how close our estimate is to the correct value θ. As another example, we might have to decide whether the unknown proportion P of patients in the imipramine group (Example 7.1.3) is larger or smaller than some specified constant when the consequences depend on where P lies relative to the constant. This last type of inference is closely related to *hypothesis testing*, the subject of Chapter 9.

Experimental Design In some statistical inference problems, we have some control over the type or the amount of experimental data that will be collected. For example, consider an experiment to determine the mean tensile strength of a certain type of alloy as a function of the pressure and temperature at which the alloy is produced. Within the limits of certain budgetary and time constraints, it may be possible for the experimenter to choose the levels of pressure and temperature at which experimental specimens of the alloy are to be produced, and also to specify the number of specimens to be produced at each of these levels.

Such a problem, in which the experimenter can choose (at least to some extent) the particular experiment that is to be carried out, is called a problem of *experimental design*. Of course, the design of an experiment and the statistical analysis of the experimental data are closely related. One cannot design an effective experiment without considering the subsequent statistical analysis that is to be carried out on the data that will be obtained. And one cannot carry out a meaningful statistical analysis of experimental data without considering the particular type of experiment from which the data were derived.

Other Inferences The general classes of problems described above, as well as the more specific examples that appeared earlier, are intended as illustrations of types of statistical inferences that we will be able to perform with the theory and methods introduced in this text. The range of possible models, inferences, and methods that can arise when data are observed in real research problems far exceeds what we can introduce here. It is hoped that gaining an understanding of the problems that we can cover here will give the reader an appreciation for what needs to be done when a more challenging statistical problem arises.

Definition of a Statistic

Example 7.1.9 Failure Times of Ball Bearings. In Example 5.6.9, we had a sample of the numbers of millions of revolutions before failure for 23 ball bearings. We modeled the lifetimes as a random sample from a lognormal distribution. We might suppose that the parameters μ and σ^2 of that lognormal distribution are unknown and that we might wish to make some inference about them. We would want to make use of the 23 observed values in making any such inference. But do we need to keep track of all 23 values or are there some summaries of the data on which our inference will be based? ◄

Each statistical inference that we will learn how to perform in this book will be based on one or a few summaries of the available data. Such data summaries arise so often and are so fundamental to inference that they receive a special name.

Definition 7.1.4 Statistic. Suppose that the observable random variables of interest are $X_1, \dots, X_n$. Let r be an arbitrary real-valued function of n real variables. Then the random variable $T = r(X_1, \dots, X_n)$ is called a *statistic*.

Three examples of statistics are the sample mean $\overline{X}_n$, the maximum Y_n of the values of $X_1, \dots, X_n$, and the function $r(X_1, \dots, X_n)$, which has the constant value 3 for all values of $X_1, \dots, X_n$.

Example 7.1.10 Failure Times of Ball Bearings. In Example 7.1.9, suppose that we were interested in making a statement about how far μ is from 40. Then we might want to use the statistic

$$T = \left| \frac{1}{36} \sum_{i=1}^{36} \log(X_i) - 4 \right|$$

in our inference procedure. In this case, T is a naïve measure of how far the data suggest that μ is from 40. ◂

Example 7.1.11 Interval for Mean. In Example 7.1.6, we constructed an interval that has probability 0.95 of containing μ. The endpoints of that interval, namely, $\overline{X}_n - 0.98$ and $\overline{X}_n + 0.98$, are statistics. ◂

Many inferences can proceed without explicitly constructing statistics as a preliminary step. However, most inferences will involve the use of statistics that could be identified in advance. And knowing which statistics are useful in which inferences can greatly simplify the implementation of the inference. Expressing an inference in terms of statistics can also help us to decide how well the inference meets out needs. For instance, in Example 7.1.10, if we estimate $|\mu - 40|$ by T, we can use the distribution of T to help determine how likely it is that T differs from $|\mu - 40|$ by a large amount. As we construct specific inferences later in this book, we will draw attention to those statistics that play important roles in the inference.

Parameters as Random Variables

There is some controversy over whether parameters should be treated as random variables or merely as numbers that index a distribution. For instance, in Example 7.1.3, we let P stand for the proportion of the patients who avoid relapse from a large group receiving imipramine. We then say that $X_1, X_2, \dots$ are i.i.d. Bernoulli random variables with parameter p conditional on $P = p$. Here, we are explicitly thinking of P as a random variable, and we give it a distribution. An alternative would be to say that $X_1, X_2, \dots$ are i.i.d. Bernoulli random variables with parameter p where p is unknown and leave it at that.

If we really want to compute something like the conditional probability that the proportion P is greater than 0.5 given the observations of the first 40 patients, then we need the conditional distribution of P given the first 40 patients, and we must treat P as a random variable. On the other hand, if we are only interested in making probability statements that are indexed by the value of p, then we do not need to think about a random variable called P. For example, we might wish to find two random variables Y_1 and Y_2 (functions of $X_1, \dots, X_{40}$) such that, no matter what p

equals, the probability that $Y_1 \leq p \leq Y_2$ is at least 0.9. Some of the inferences that we shall discuss later in this book are of the former type that require treating P as a random variable, and some are of the latter type in which p is merely an index for a distribution.

Some statisticians believe that it is possible and useful to treat parameters as random variables in every statistical inference problem. They believe that the distribution of the parameter is a subjective probability distribution in the sense that it represents an individual experimenter's information and subjective beliefs about where the true value of the parameter is likely to lie. Once they assign a distribution for a parameter, that distribution is no different from any other probability distribution used in the field of statistics, and all of the rules of probability theory apply to every distribution. Indeed, in all of the cases described in this book, the parameters can actually be identified as limits of functions of large collections of potential observations. Here is a typical example.

Example 7.1.12 Parameter as a Limit of Random Variables. In Example 7.1.3, the parameter P can be understood as follows: Imagine an infinite sequence of potential patients receiving imipramine treatment. Assume that for every integer n, the outcomes of every ordered subset of n patients from that infinite sequence has the same joint distribution as the outcomes of every other ordered subset of n patients. In other words, assume that the order in which the patients appear in the sequence is irrelevant to the joint distribution of the patient outcomes. Let P_n be the proportion of patients who don't relapse out of the first n. It can be shown that the probability is 1 that P_n converges to something as $n \to \infty$. That something can be thought of as P, which we have been calling the proportion of successes in a very large population. In this sense, P is a random variable because it is a function of other random variables. A similar argument can be made in all of the statistical models in this book involving parameters, but the mathematics needed to make these arguments precise is too advanced to present here. (Chapter 1 of Schervish (1995) contains the necessary details.) Statisticians who argue as in this example are said to adhere to the Bayesian philosophy of statistics and are called *Bayesians*. ◀

There is another line of reasoning that leads naturally to treating P as a random variable in Example 7.1.12 without relying on an infinite sequence of potential patients. Suppose that the number of potential patients is enough larger than any sample that we will see to make the approximation in Theorem 5.3.4 applicable. Then P is just the proportion of successes among the large population of potential patients. Conditional on $P = p$, the number of successes in a sample of n patients will be approximately a binomial random variable with paramters n and p according to Theorem 5.3.4. If the outcomes of the patients in the sample are random variables, then it makes sense that the proportion of successes among those patients is also random.

There is another group of statisticians who believe that in many problems it is not appropriate to assign a distribution to a parameter but claim instead that the true value of the parameter is a certain fixed number whose value happens to be unknown to the experimenter. These statisticians would assign a distribution to a parameter only when there is extensive previous information about the relative frequencies with which similar parameters have taken each of their possible values in past experiments. If two different scientists could agree on which past experiments were similar to the present experiment, then they might agree on a distribution to be assigned to the parameter. For example, suppose that the proportion θ of defective items in a certain large manufactured lot is unknown. Suppose also that

the same manufacturer has produced many such lots of items in the past and that detailed records have been kept about the proportions of defective items in past lots. The relative frequencies for past lots could then be used to construct a distribution for θ. Statisticians who would argue this way are said to adhere to the frequentist philosophy of statistics and are called *frequentists*.

The frequentists rely on the assumption that there exist infinite sequences of random variables in order to make sense of most of their probability statements. Once one assumes the existence of such an infinite sequence, one finds that the parameters of the distributions being used are limits of functions of the infinite sequences, just as do the Bayesians described above. In this way, the parameters are random variables because they are functions of random variables. The point of disagreement between the two groups is whether it is useful or even possible to assign a distribution to such parameters.

Both Bayesians and frequentists agree on the usefulness of families of distributions for observations indexed by parameters. Bayesians refer to the distribution indexed by parameter value θ as the conditional distribution of the observations given that the parameter equals θ. Frequentists refer to the distribution indexed by θ as the distribution of the observations when θ is the true value of the parameter. The two groups agree that whenever a distribution can be assigned to a parameter, the theory and methods to be described in this chapter are applicable and useful. In Sections 7.2–7.4, we shall explicitly assume that each parameter is a random random variable and we shall assign it a distribution that represents the probabilities that the parameter lies in various subsets of the parameter space. Beginning in Sec. 7.5, we shall consider techniques of estimation that are not based on assigning distributions to parameters.

References

In the remainder of this book, we shall consider many different problems of statistical inference, statistical decision, and experimental design. Some books that discuss statistical theory and methods at about the same level as they will be discussed in this book were mentioned at the end of Sec. 1.1. Some statistics books written at a more advanced level are Bickel and Doksum (2000), Casella and Berger (2002), Cramér (1946), DeGroot (1970), Ferguson (1967), Lehmann (1997), Lehmann and Casella (1998), Rao (1973), Rohatgi (1976), and Schervish (1995).

Exercises

1. Identify the components of the statistical model (as defined in Definition 7.1.1) in Example 7.1.3.

2. Identify two statistical inferences mentioned in Example 7.1.3.

3. In Examples 7.1.4 and 5.7.8 (page 323), identify the components of the statistical model as defined in Definition 7.1.1.

4. In Example 7.1.6, identify the components of the statistical model as defined in Definition 7.1.1.

5. In Example 7.1.6, identify any statistical inference mentioned.

6. In Example 5.8.3 (page 328), identify the components of the statistical model as defined in Definition 7.1.1.

7. In Example 5.4.7 (page 293), identify the components of the statistical model as defined in Definition 7.1.1.

7.2 Prior and Posterior Distributions

The distribution of a parameter before observing any data is called the prior distribution of the parameter. The conditional distribution of the parameter given the observed data is called the posterior distribution. If we plug the observed values of the data into the conditional p.f. or p.d.f. of the data given the parameter, the result is a function of the parameter alone, which is called the likelihood function.

The Prior Distribution

Example 7.2.1 Lifetimes of Electronic Components. In Example 7.1.1, lifetimes $X_1, X_2, \ldots$ of electronic components were modeled as i.i.d. exponential random variables with parameter θ conditional on θ, and θ was interpreted as the failure rate of the components. Indeed, we noted that $n/\sum_{i=1}^{n} X_i$ should converge in probability to θ as n goes to ∞. We then said that θ had the gamma distribution with parameters 1 and 2. ◀

The distribution of θ mentioned at the end of Example 7.2.1 was assigned before observing any of the component lifetimes. For this reason, we call it a *prior distribution*.

Definition 7.2.1 Prior Distribution/p.f./p.d.f. Suppose that one has a statistical model with parameter θ. If one treats θ as random, then the distribution that one assigns to θ before observing the other random variables of interest is called its *prior distribution*. If the parameter space is at most countable, then the prior distribution is discrete and its p.f. is called the *prior p.f.* of θ. If the prior distribution is a continuous distribution, then its p.d.f. is called the *prior p.d.f.* of θ. We shall commonly use the symbol $\xi(\theta)$ to denote the prior p.f. or p.d.f. as a function of θ.

When one treats the parameter as a random variable, the name "prior distribution" is merely another name for the marginal distribution of the parameter.

Example 7.2.2 Fair or Two-Headed Coin. Let θ denote the probability of obtaining a head when a certain coin is tossed, and suppose that it is known that the coin either is fair or has a head on each side. Therefore, the only possible values of θ are $\theta = 1/2$ and $\theta = 1$. If the prior probability that the coin is fair is 0.8, then the prior p.f. of θ is $\xi(1/2) = 0.8$ and $\xi(1) = 0.2$. ◀

Example 7.2.3 Proportion of Defective Items. Suppose that the proportion θ of defective items in a large manufactured lot is unknown and that the prior distribution assigned to θ is the uniform distribution on the interval [0, 1]. Then the prior p.d.f. of θ is

$$\xi(\theta) = \begin{cases} 1 & \text{for } 0 < \theta < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{7.2.1}$$

◀

The prior distribution of a parameter θ must be a probability distribution over the parameter space Ω. We assume that the experimenter or statistician will be able to summarize his previous information and knowledge about where in Ω the value of θ is likely to lie by constructing a probability distribution on the set Ω. In other words, before the experimental data have been collected or observed, the experimenter's past experience and knowledge will lead him to believe that θ is more likely to lie in certain regions of Ω than in others. We shall assume that the relative likelihoods

of the different regions can be expressed in terms of a probability distribution on Ω, namely, the prior distribution of θ.

Example 7.2.4 Lifetimes of Fluorescent Lamps. Suppose that the lifetimes (in hours) of fluorescent lamps of a certain type are to be observed and that the the lifetime of any particular lamp has the exponential distribution with parameter θ. Suppose also that the exact value of θ is unknown, and on the basis of previous experience the prior distribution of θ is taken as the gamma distribution for which the mean is 0.0002 and the standard deviation is 0.0001. We shall determine the prior p.d.f. of θ.

Suppose that the prior distribution of θ is the gamma distribution with parameters α_0 and β_0. It was shown in Theorem 5.7.5 that the mean of this distribution is α_0/β_0 and the variance is α_0/β_0^2. Therefore, $\alpha_0/\beta_0 = 0.0002$ and $\alpha_0^{1/2}/\beta_0 = 0.0001$. Solving these two equations gives $\alpha_0 = 4$ and $\beta_0 = 20{,}000$. It follows from Eq. (5.7.13) that the prior p.d.f. of θ for $\theta > 0$ is as follows:

$$\xi(\theta) = \frac{(20{,}000)^4}{3!}\theta^3 e^{-20{,}000\theta}. \tag{7.2.2}$$

Also, $\xi(\theta) = 0$ for $\theta \le 0$. ◂

In the remainder of this section and Sections 7.3 and 7.4, we shall focus on statistical inference problems in which the parameter θ is a random variable of interest and hence will need to be assigned a distribution. In such problems, we shall refer to the distribution indexed by θ for the other random variables of interest as the conditional distribution for those random variables given θ. For example, this is precisely the language used in Example 7.2.1 where the parameter is θ, the failure rate. In referring to the conditional p.f. or p.d.f. of random variables, such as $X_1, X_2, \ldots$ in Example 7.2.1, we shall use the notation of conditional p.f.'s and p.d.f.'s. For example, if we let $\boldsymbol{X} = (X_1, \ldots, X_m)$ in Example 7.2.1, the conditional p.d.f. of $\boldsymbol{X}$ given θ is

$$f_m(\boldsymbol{x}|\theta) = \begin{cases} \theta^m \exp(-\theta[x_1 + \cdots + x_m]) & \text{for all } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{7.2.3}$$

In many problems, such as Example 7.2.1, the observable data $X_1, X_2, \ldots$ are modeled as a random sample from a univariate distribution indexed by θ. In these cases, let $f(x|\theta)$ denote the p.f. or p.d.f. of a single random variable under the distribution indexed by θ. In such a case, using the above notation,

$$f_m(\boldsymbol{x}|\theta) = f(x_1|\theta) \cdots f(x_m|\theta).$$

When we treat θ as a random variable, $f(x|\theta)$ is the conditional p.f. or p.d.f. of each observation X_i given θ, and the observations are conditionally i.i.d. given θ. In summary, the following two expressions are to be understood as equivalent:

- $X_1, \ldots, X_n$ form a random sample with p.f. or p.d.f. $f(x|\theta)$.
- $X_1, \ldots, X_n$ are conditionally i.i.d. given θ with conditional p.f. or p.d.f. $f(x|\theta)$.

Although we shall generally use the wording in the first bullet above for simplicity, it is often useful to remember that the two wordings are equivalent when we treat θ as a random variable.

Sensitivity Analysis and Improper Priors In Example 2.3.8 on page 84, we saw a situation in which two very different sets of prior probabilities were used for a collection of events. After we observed data, however, the posterior probabilities were

quite similar. In Example 5.8.4 on page 330, we used a large collection of prior distributions for a parameter in order to see how much impact the prior distribution had on the posterior probability of a single important event. It is a common practice to compare the posterior distributions that arise from several different prior distributions in order to see how much effect the prior distribution has on the answers to important questions. Such comparisons are called *sensitivity analysis*.

It is very often the case that different prior distributions do not make much difference after the data have been observed. This is especially true if there are a lot of data or if the prior distributions being compared are very spread out. This observation has two important implications. First, the fact that different experimenters might not agree on a prior distribution becomes less important if there are a lot of data. Second, experimenters might be less inclined to spend time specifying a prior distribution if it is not going to matter much which one is specified. Unfortunately, if one does not specify some prior distribution, there is no way to calculate a conditional distribution of the parameter given the data.

As an expedient, there are some calculations available that attempt to capture the idea that the data contain much more information than is available a priori. Usually, these calculations involve using a function $\xi(\theta)$ as if it were a prior p.d.f. for the parameter θ but such that $\int \xi(\theta)\, d\theta = \infty$, which clearly violates the definition of p.d.f. Such priors are called *improper*. We shall discuss improper priors in more detail in Sec. 7.3.

The Posterior Distribution

Example 7.2.5 Lifetimes of Fluorescent Lamps. In Example 7.2.4, we constructed a prior distribution for the parameter θ that specifies the exponential distribution for a collection of lifetimes of fluorescent lamps. Suppose that we observe a collection of n such lifetimes. How would we change the distribution of θ to take account of the observed data? ◀

Definition 7.2.2 Posterior Distribution/p.f./p.d.f. Consider a statistical inference problem with parameter θ and random variables $X_1, \dots, X_n$ to be observed. The conditional distribution of θ given $X_1, \dots, X_n$ is called the *posterior distribution* of θ. The conditional p.f. or p.d.f. of θ given $X_1 = x_1, \dots, X_n = x_n$ is called the *posterior p.f.* or *posterior p.d.f.* of θ and is typically denoted $\xi(\theta|x_1, \dots, x_n)$.

When one treats the parameter as a random variable, the name "posterior distribution" is merely another name for the conditional distribution of the parameter given the data. Bayes' theorem for random variables (3.6.13) and for random vectors (3.7.15) tells us how to compute the posterior p.d.f. or p.f. of θ after observing data. We shall review the derivation of Bayes' theorem here using the specific notation of prior distributions and parameters.

Theorem 7.2.1 Suppose that the n random variables $X_1, \dots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$. Suppose also that the value of the parameter θ is unknown and the prior p.d.f. or p.f. of θ is $\xi(\theta)$. Then the posterior p.d.f. or p.f. of θ is

$$\xi(\theta|\boldsymbol{x}) = \frac{f(x_1|\theta)\cdots f(x_n|\theta)\xi(\theta)}{g_n(\boldsymbol{x})} \quad \text{for } \theta \in \Omega,$$

where g_n is the marginal joint p.d.f. or p.f. of $X_1, \dots, X_n$.

Proof For simplicity, we shall assume that the parameter space Ω is either an interval of the real line or the entire real line and that $\xi(\theta)$ is a prior p.d.f. on Ω, rather than a prior p.f. However, the proof that will be given here can be adapted easily to a problem in which $\xi(\theta)$ is a p.f.

Since the random variables $X_1, \ldots, X_n$ form a random sample from the distribution for which the p.d.f. is $f(x|\theta)$, it follows from Sec. 3.7 that their conditional joint p.d.f. or p.f. $f_n(x_1, \ldots, x_n|\theta)$ given θ is

$$f_n(x_1, \ldots, x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta). \tag{7.2.4}$$

If we use the vector notation $\boldsymbol{x} = (x_1, \ldots, x_n)$, then the joint p.d.f. in Eq. (7.2.4) can be written more compactly as $f_n(\boldsymbol{x}|\theta)$. Eq. (7.2.4) merely expresses the fact that $X_1, \ldots, X_n$ are conditionally independent and identically distributed given θ, each having p.d.f. or p.f. $f(x|\theta)$.

If we multiply the conditional joint p.d.f. or p.f. by the p.d.f. $\xi(\theta)$, we obtain the $(n+1)$-dimensional joint p.d.f. (or p.f./p.d.f.) of $X_1, \ldots, X_n$ and θ in the form

$$f(\boldsymbol{x}, \theta) = f_n(\boldsymbol{x}|\theta)\xi(\theta). \tag{7.2.5}$$

The marginal joint p.d.f. or p.f. of $X_1, \ldots, X_n$ can now be obtained by integrating the right-hand side of Eq. (7.2.5) over all values of θ. Therefore, the n-dimensional marginal joint p.d.f. or p.f. $g_n(\boldsymbol{x})$ of $X_1, \ldots, X_n$ can be written in the form

$$g_n(\boldsymbol{x}) = \int_{\Omega} f_n(\boldsymbol{x}|\theta)\xi(\theta)\, d\theta. \tag{7.2.6}$$

Eq. (7.2.6) is just an instance of the law of total probability for random vectors (3.7.14).

Furthermore, the conditional p.d.f. of θ given that $X_1 = x_1, \ldots, X_n = x_n$, namely, $\xi(\theta|\boldsymbol{x})$, must be equal to $f(\boldsymbol{x}, \theta)$ divided by $g_n(\boldsymbol{x})$. Thus, we have

$$\xi(\theta|\boldsymbol{x}) = \frac{f_n(\boldsymbol{x}|\theta)\xi(\theta)}{g_n(\boldsymbol{x})} \quad \text{for } \theta \in \Omega, \tag{7.2.7}$$

which is Bayes' theorem restated for parameters and random samples. If $\xi(\theta)$ is a p.f., so that the prior distribution is discrete, just replace the integral in (7.2.6) by the sum over all of the possible values of θ. ■

Example 7.2.6 Lifetimes of Fluorescent Lamps. Suppose again, as in Examples 7.2.4 and 7.2.5, that the distribution of the lifetimes of fluorescent lamps of a certain type is the exponential distribution with parameter θ, and the prior distribution of θ is a particular gamma distribution for which the p.d.f. $\xi(\theta)$ is given by Eq. (7.2.2). Suppose also that the lifetimes $X_1, \ldots, X_n$ of a random sample of n lamps of this type are observed. We shall determine the posterior p.d.f. of θ given that $X_1 = x_1, \ldots, X_n = x_n$.

By Eq. (5.7.16), the p.d.f. of each observation X_i is

$$f(x|\theta) = \begin{cases} \theta e^{-\theta x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The joint p.d.f. of $X_1, \ldots, X_n$ can be written in the following form, for $x_i > 0$ ($i = 1, \ldots, n$):

$$f_n(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta y},$$

where $y = \sum_{i=1}^{n} x_i$. As $f_n(\boldsymbol{x}|\theta)$ will be used in constructing the posterior distribution of θ, it is now apparent that the statistic $Y = \sum_{i=1}^{n} X_i$ will be used in any inference that makes use of the posterior distribution.

Since the prior p.d.f. $\xi(\theta)$ is given by Eq. (7.2.2), it follows that for $\theta > 0$,

$$f_n(\boldsymbol{x}|\theta)\xi(\theta) = \theta^{n+3}e^{-(y+20{,}000)\theta}. \tag{7.2.8}$$

We need to compute $g_n(\boldsymbol{x})$, which is the integral of (7.2.8) over all θ:

$$g_n(\boldsymbol{x}) = \int_0^\infty \theta^{n+3}e^{-(y+20{,}000)\theta}d\theta = \frac{\Gamma(n+4)}{(y+20{,}000)^{n+4}},$$

where the last equality follows from Theorem 5.7.3. Hence,

$$\begin{aligned}\xi(\theta|\boldsymbol{x}) &= \frac{\theta^{n+3}e^{-(y+20{,}000)\theta}}{\dfrac{\Gamma(n+4)}{(y+20{,}000)^{n+4}}} \\ &= \frac{(y+20{,}000)^{n+4}}{\Gamma(n+4)}e^{-(y+20{,}000)\theta},\end{aligned} \tag{7.2.9}$$

for $\theta > 0$. When we compare this expression with Eq. (5.7.13), we can see that it is the p.d.f. of the gamma distribution with parameters $n+4$ and $y+20{,}000$. Hence, this gamma distribution is the posterior distribution of θ.

As a specific example, suppose that we observe the following $n = 5$ lifetimes in hours: 2911, 3403, 3237, 3509, and 3118. Then $y = 16{,}178$, and the posterior distribution of θ is the gamma distribution with parameters 9 and 36,178. The top panel of Fig. 7.1 displays both the prior and posterior p.d.f.'s in this example. It is clear that the data have caused the distribution of θ to change somewhat from the prior to the posterior.

At this point, it might be appropriate to perform a sensitivity analysis. For example, how would the posterior distribution change if we had chosen a different prior distribution? To be specific, consider the gamma prior with parameters 1 and 1000. This prior has the same standard deviation as the original prior, but the mean is five times as big. The posterior distribution would then be the gamma distribution with parameters 6 and 17,178. The p.d.f.'s of this pair of prior and posterior are plotted in the lower panel of Fig. 7.1. One can see that both the prior and the posterior in the bottom panel are more spread out than their counterparts in the upper panel. It

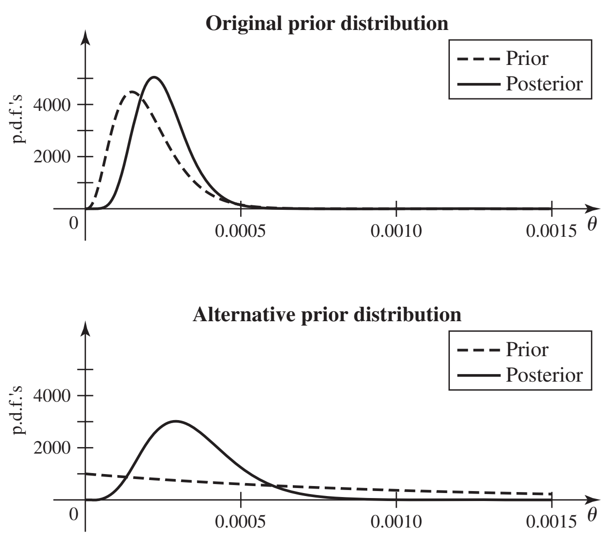


Figure 7.1 Prior and posterior p.d.f.'s in Example 7.2.6. The top panel is based on the original prior. The bottom panel is based on the alternative prior that was part of the sensitivity analysis.

is clear that the choice of prior distribution is going to make a difference with this small data set. ◀

The names "prior" and "posterior" derive from the Latin words for "former" and "coming after." The prior distribution is the distribution of θ that comes before observing the data, and posterior distribution comes after observing the data.

The Likelihood Function

The denominator on the right side of Eq. (7.2.7) is simply the integral of the numerator over all possible values of θ. Although the value of this integral depends on the observed values $x_1, \ldots, x_n$, it does not depend on θ and it may be treated as a constant when the right-hand side of Eq. (7.2.7) is regarded as a p.d.f. of θ. We may therefore replace Eq. (7.2.7) with the following relation:

$$\xi(\theta|\boldsymbol{x}) \propto f_n(\boldsymbol{x}|\theta)\xi(\theta). \tag{7.2.10}$$

The proportionality symbol $\propto$ is used here to indicate that the left side is equal to the right side except possibly for a constant factor, the value of which may depend on the observed values $x_1, \ldots, x_n$ but does not depend on θ. The appropriate constant factor that will establish the equality of the two sides in the relation (7.2.10) can be determined at any time by using the fact that $\int_\Omega \xi(\theta|\boldsymbol{x})\, d\theta = 1$, because $\xi(\theta|\boldsymbol{x})$ is a p.d.f. of θ.

One of the two functions on the right-hand side of Eq. (7.2.10) is the prior p.d.f. of θ. The other function has a special name also.

Definition 7.2.3 Likelihood Function. When the joint p.d.f. or the joint p.f. $f_n(\boldsymbol{x}|\theta)$ of the observations in a random sample is regarded as a function of θ for given values of $x_1, \ldots, x_n$, it is called the *likelihood function*.

The relation (7.2.10) states that the posterior p.d.f. of θ is proportional to the product of the likelihood function and the prior p.d.f. of θ.

By using the proportionality relation (7.2.10), it is often possible to determine the posterior p.d.f. of θ without explicitly performing the integration in Eq. (7.2.6). If we can recognize the right side of the relation (7.2.10) as being equal to one of the standard p.d.f.'s introduced in Chapter 5 or elsewhere in this book, except possibly for a constant factor, then we can easily determine the appropriate factor that will convert the right side of (7.2.10) into a proper p.d.f. of θ. We shall illustrate these ideas by considering again Example 7.2.3.

Example 7.2.7 Proportion of Defective Items. Suppose again, as in Example 7.2.3, that the proportion θ of defective items in a large manufactured lot is unknown and that the prior distribution of θ is a uniform distribution on the interval [0, 1]. Suppose also that a random sample of n items is taken from the lot, and for $i = 1, \ldots, n$, let $X_i = 1$ if the ith item is defective, and let $X_i = 0$ otherwise. Then $X_1, \ldots, X_n$ form n Bernoulli trials with parameter θ. We shall determine the posterior p.d.f. of θ.

It follows from Eq. (5.2.2) that the p.f. of each observation X_i is

$$f(x|\theta) = \begin{cases} \theta^x(1-\theta)^{1-x} & \text{for } x = 0, 1, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, if we let $y = \sum_{i=1}^n x_i$, then the joint p.f. of $X_1, \ldots, X_n$ can be written in the following form for $x_i = 0$ or 1 ($i = 1, \ldots, n$):

$$f_n(\boldsymbol{x}|\theta) = \theta^y(1-\theta)^{n-y}. \tag{7.2.11}$$

Since the prior p.d.f. $\xi(\theta)$ is given by Eq. (7.2.1), it follows that for $0 < \theta < 1$,

$$f_n(\boldsymbol{x}|\theta)\xi(\theta) = \theta^y(1-\theta)^{n-y}. \tag{7.2.12}$$

When we compare this expression with Eq. (5.8.3), we can see that, except for a constant factor, it is the p.d.f. of the beta distribution with parameters $\alpha = y + 1$ and $\beta = n - y + 1$. Since the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ is proportional to the right side of Eq. (7.2.12), it follows that $\xi(\theta|\boldsymbol{x})$ must be the p.d.f. of the beta distribution with parameters $\alpha = y + 1$ and $\beta = n - y + 1$. Therefore, for $0 < \theta < 1$,

$$\xi(\theta|\boldsymbol{x}) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y}. \tag{7.2.13}$$

In this example, the statistic $Y = \sum_{i=1}^{n} X_i$ is being used to construct the posterior distribution, and hence will be used in any inference that is based on the posterior distribution. ◀

Note: Normalizing Constant for Posterior p.d.f. The steps that got us from (7.2.12) to (7.2.13) are an example of a very common technique for determining a posterior p.d.f. We can drop any inconvenient constant factor from the prior p.d.f. and from the likelihood function before we multiply them together as in (7.2.10). Then we look at the resulting product, call it $g(\theta)$, to see if we recognize it as looking like part of a p.d.f. that we have seen elsewhere. If indeed we find a named distribution with p.d.f. equal to $cg(\theta)$, then our posterior p.d.f. is also $cg(\theta)$, and our posterior distribution has the corresponding name, just as in Example 7.2.7.

Sequential Observations and Prediction

In many experiments, the observations $X_1, \ldots, X_n$, which form the random sample, must be obtained sequentially, that is, one at a time. In such an experiment, the value of X_1 is observed first, the value of X_2 is observed next, the value of X_3 is then observed, and so on. Suppose that the prior p.d.f. of the parameter θ is $\xi(\theta)$. After the value x_1 of X_1 has been observed, the posterior p.d.f. $\xi(\theta|x_1)$ can be calculated in the usual way from the relation

$$\xi(\theta|x_1) \propto f(x_1|\theta)\xi(\theta). \tag{7.2.14}$$

Since X_1 and X_2 are conditionally independent given θ, the conditional p.f. or p.d.f. of X_2 given θ and $X_1 = x_1$ is the same as that given θ alone, namely, $f(x_2|\theta)$. Hence, the posterior p.d.f. of θ in Eq. (7.2.14) serves as the prior p.d.f. of θ when the value of X_2 is to be observed. Thus, after the value x_2 of X_2 has been observed, the posterior p.d.f. $\xi(\theta|x_1, x_2)$ can be calculated from the relation

$$\xi(\theta|x_1, x_2) \propto f(x_2|\theta)\xi(\theta|x_1). \tag{7.2.15}$$

We can continue in this way, calculating an updated posterior p.d.f. of θ after each observation and using that p.d.f. as the prior p.d.f. of θ for the next observation. The posterior p.d.f. $\xi(\theta|x_1, \ldots, x_{n-1})$ after the values $x_1, \ldots, x_{n-1}$ have been observed will ultimately be the prior p.d.f. of θ for the final observed value of X_n. The posterior p.d.f. after all n values $x_1, \ldots, x_n$ have been observed will therefore be specified by the relation

$$\xi(\theta|\boldsymbol{x}) \propto f(x_n|\theta)\xi(\theta|x_1, \ldots, x_{n-1}). \tag{7.2.16}$$

Alternatively, after all n values $x_1, \ldots, x_n$ have been observed, we could calculate the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ in the usual way by combining the joint p.d.f. $f_n(\boldsymbol{x}|\theta)$ with the original prior p.d.f. $\xi(\theta)$, as indicated in Eq. (7.2.7). It can be shown (see

Exercise 8) that the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ will be the same regardless of whether it is calculated directly by using Eq. (7.2.7) or sequentially by using Eqs. (7.2.14), (7.2.15), and (7.2.16). This property was illustrated in Sec. 2.3 (see page 80) for a coin that is known either to be fair or to have a head on each side. After each toss of the coin, the posterior probability that the coin is fair is updated.

The proportionality constants in Eqs. (7.2.14)–(7.2.16) have a useful interpretation. For example, in (7.2.16) the proportionality constant is 1 over the integral of the right side with respect to θ. But this integral is the conditional p.d.f. or p.f. of X_n given $X_1 = x_1, \dots, X_{n-1} = x_{n-1}$, according to the conditional version of the law of total probability (3.7.16). For example, if θ has a continuous distribution,

$$f(x_n|x_1, \dots, x_{n-1}) = \int f(x_n|\theta)\xi(\theta|x_1, \dots, x_{n-1})d\theta. \tag{7.2.17}$$

The proportionality constant in (7.2.16) is 1 over (7.2.17). So, if we are interested in predicting the nth observation in a sequence after observing the first $n-1$, we can use (7.2.17), which is also 1 over the proportionality constant in Eq. (7.2.16), as the conditional p.f. or p.d.f. of X_n given the first $n-1$ observations.

Example 7.2.8

Lifetimes of Fluorescent Lamps. In Example 7.2.6, conditional on θ, the lifetimes of fluorescent lamps are independent exponential random variables with parameter θ. We also observed the lifetimes of five lamps, and the posterior distribution of θ was found to be the gamma distribution with parameters 9 and 36,178. Suppose that we want to predict the lifetime X_6 of the next lamp.

The conditional p.d.f. of X_6, the lifetime of the next lamp, given the first five lifetimes equals the integral of $\xi(\theta|\boldsymbol{x})f(x_6|\theta)$ with respect to θ. The posterior p.d.f. of θ is $\xi(\theta|\boldsymbol{x}) = 2.633 \times 10^{36}\theta^8 e^{-36,178\theta}$ for $\theta > 0$. So, for $x_6 > 0$

$$\begin{aligned} f(x_6|\boldsymbol{x}) &= \int_0^\infty 2.633 \times 10^{36}\theta^8 e^{-36,178\theta}\theta e^{-x_6\theta}\, d\theta \\ &= 2.633 \times 10^{36}\int_0^\infty \theta^9 e^{-(x_6+36,178)\theta}\, d\theta \\ &= 2.633 \times 10^{36}\frac{\Gamma(10)}{(x_6+36,178)^{10}} = \frac{9.555 \times 10^{41}}{(x_6+36,178)^{10}}. \end{aligned} \tag{7.2.18}$$

We can use this p.d.f. to perform any calculation we wish concerning the distribution of X_6 given the observed lifetimes. For example, the probability that the sixth lamp lasts more than 3000 hours equals

$$\Pr(X_6 > 3000|\boldsymbol{x}) = \int_{3000}^\infty \frac{9.555 \times 10^{41}}{(x_6+36,178)^{10}}\,dx_6 = \frac{9.555 \times 10^{41}}{9 \times 39,178^9} = 0.4882.$$

Finally, we can continue the sensitivity analysis that was started in Example 7.2.6. If it is important to know the probability that the next lifetime is at least 3000, we can see how much influence the choice of prior distribution has made on this calculation. Using the second prior distribution (gamma with parameters 1 and 1000), we found that the posterior distribution of θ was the gamma distribution with parameters 6 and 17,178. We could compute the conditional p.d.f. of X_6 given the observed data in the same way as we did with the original posterior, and it would be

$$f(x_6|\boldsymbol{x}) = \frac{1.542 \times 10^{26}}{(x_6+17,178)^7}, \quad \text{for } x_6 > 0. \tag{7.2.19}$$

With this p.d.f., the probability that $X_6 > 3000$ is

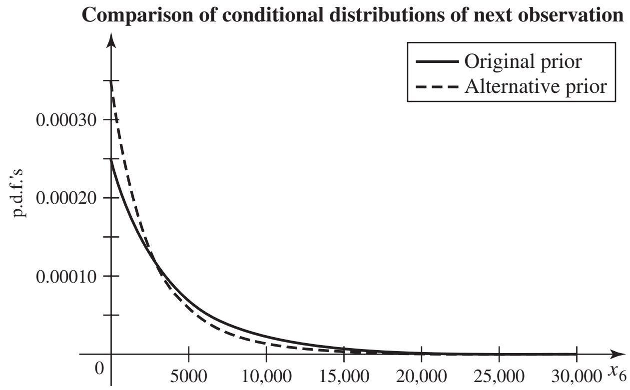


Figure 7.2 Two possible conditional p.d.f.'s, Eqs. (7.2.18) and (7.2.19) for X_6 given the observed data in Example 7.2.8. The two p.d.f.'s were computed using the two different posterior distributions that were derived from the two different prior distributions in Example 7.2.6.

$$\Pr(X_6 > 3000|\boldsymbol{x}) = \int_{3000}^{\infty} \frac{1.542 \times 10^{26}}{(x_6 + 17{,}178)^7}\,dx_6 = \frac{1.542 \times 10^{26}}{6 \times 20{,}178^6} = 0.3807.$$

As we noted at the end of Example 7.2.6, the different priors make a considerable difference in the inferences that we can make. If it is important to have a precise value of $\Pr(X_6 > 3000|\boldsymbol{x})$, we need a larger sample. The two different p.d.f.'s of X_6 given $\boldsymbol{x}$ can be compared in Fig. 7.2. The p.d.f. from Eq. (7.2.18) is higher for intermediate values of x_6, while the one from Eq. (7.2.19) is higher for the extreme values of x_6. ◄

Summary

The prior distribution of a parameter describes our uncertainty about the parameter before observing any data. The likelihood function is the conditional p.d.f. or p.f. of the data given the parameter when regarded as a function of the parameter with the observed data plugged in. The likelihood tells us how much the data will alter our uncertainty. Large values of the likelihood correspond to parameter values where the posterior p.d.f. or p.f. will be higher than the prior. Low values of the likelihood occur at parameter values where the posterior will be lower than the prior. The posterior distribution of the parameter is the conditional distribution of the parameter given the data. It is obtained using Bayes' theorem for random variables, which we first saw on page 148. We can predict future observations that are conditionally independent of the observed data given θ by using the conditional version of the law of total probability that we saw on page 163.

Exercises

1. Consider again the situation described in Example 7.2.8. This time, suppose that the experimenter believes that the prior distribution of θ is the gamma distribution with parameters 1 and 5000. What would this experimenter compute as the value of $\Pr(X_6 > 3000|\boldsymbol{x})$?

2. Suppose that the proportion θ of defective items in a large manufactured lot is known to be either 0.1 or 0.2, and the prior p.f. of θ is as follows:

$$\xi(0.1) = 0.7 \quad \text{and} \quad \xi(0.2) = 0.3.$$

Suppose also that when eight items are selected at random from the lot, it is found that exactly two of them are defective. Determine the posterior p.f. of θ.

3. Suppose that the number of defects on a roll of magnetic recording tape has a Poisson distribution for which the mean λ is either 1.0 or 1.5, and the prior p.f. of λ is as

follows:

$$\xi(1.0) = 0.4 \quad \text{and} \quad \xi(1.5) = 0.6.$$

If a roll of tape selected at random is found to have three defects, what is the posterior p.f. of λ?

4. Suppose that the prior distribution of some parameter θ is a gamma distribution for which the mean is 10 and the variance is 5. Determine the prior p.d.f. of θ.

5. Suppose that the prior distribution of some parameter θ is a beta distribution for which the mean is 1/3 and the variance is 1/45. Determine the prior p.d.f. of θ.

6. Suppose that the proportion θ of defective items in a large manufactured lot is unknown, and the prior distribution of θ is the uniform distribution on the interval [0, 1]. When eight items are selected at random from the lot, it is found that exactly three of them are defective. Determine the posterior distribution of θ.

7. Consider again the problem described in Exercise 6, but suppose now that the prior p.d.f. of θ is as follows:

$$\xi(\theta) = \begin{cases} 2(1-\theta) & \text{for } 0 < \theta < 1, \\ 0 & \text{otherwise.} \end{cases}$$

As in Exercise 6, suppose that in a random sample of eight items exactly three are found to be defective. Determine the posterior distribution of θ.

8. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. is $f(x|\theta)$, the value of θ is unknown, and the prior p.d.f. of θ is $\xi(\theta)$. Show that the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ is the same regardless of whether it is calculated directly by using Eq. (7.2.7) or sequentially by using Eqs. (7.2.14), (7.2.15), and (7.2.16).

9. Consider again the problem described in Exercise 6, and assume the same prior distribution of θ. Suppose now, however, that instead of selecting a random sample of eight items from the lot, we perform the following experiment: Items from the lot are selected at random one by one until exactly three defectives have been found. If we find that we must select a total of eight items in this experiment, what is the posterior distribution of θ at the end of the experiment?

10. Suppose that a single observation X is to be taken from the uniform distribution on the interval $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, the value of θ is unknown, and the prior distribution of θ is the uniform distribution on the interval [10, 20]. If the observed value of X is 12, what is the posterior distribution of θ?

11. Consider again the conditions of Exercise 10, and assume the same prior distribution of θ. Suppose now, however, that six observations are selected at random from the uniform distribution on the interval $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, and their values are 11.0, 11.5, 11.7, 11.1, 11.4, and 10.9. Determine the posterior distribution of θ.

7.3 Conjugate Prior Distributions

For each of the most popular statistical models, there exists a family of distributions for the parameter with a very special property. If the prior distribution is chosen to be a member of that family, then the posterior distribution will also be a member of that family. Such a family of distributions is called a conjugate family. Choosing a prior distribution from a conjugate family will typically make it particularly simple to calculate the posterior distribution.

Sampling from a Bernoulli Distribution

Example 7.3.1 A Clinical Trial. In Example 5.8.5 (page 330), we were observing patients in a clinical trial. The proportion P of successful outcomes among all possible patients was a random variable for which we chose a distribution from the family of beta distributions. This choice made the calculation of the conditional distribution of P given the observed data very simple at the end of that example. Indeed, the conditional distribution of P given the data was another member of the beta family. ◄

That the result in Example 7.3.1 occurs in general is the subject of the next theorem.

Theorem 7.3.1 Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with parameter θ, which is unknown $(0 < \theta < 1)$. Suppose also that the prior distribution

of θ is the beta distribution with parameters $\alpha > 0$ and $\beta > 0$. Then the posterior distribution of θ given that $X_i = x_i$ $(i = 1, \ldots, n)$ is the beta distribution with parameters $\alpha + \sum_{i=1}^n x_i$ and $\beta + n - \sum_{i=1}^n x_i$.

Theorem 7.3.1 is just a restatement of Theorem 5.8.2 (page 329), and its proof is essentially the calculation in Example 5.8.3.

Updating the Posterior Distribution One implication of Theorem 7.3.1 is the following: Suppose that the proportion θ of defective items in a large shipment is unknown, the prior distribution of θ is the beta distribution with parameters α and β, and n items are selected one at a time at random from the shipment and inspected. Assume that the items are conditionally independent given θ. If the first item inspected is defective, the posterior distribution of θ will be the beta distribution with parameters $\alpha + 1$ and β. If the first item is nondefective, the posterior distribution will be the beta distribution with parameters α and $\beta + 1$. The process can be continued in the following way: Each time an item is inspected, the current posterior beta distribution of θ is changed to a new beta distribution in which the value of either the parameter α or the parameter β is increased by one unit. The value of α is increased by one unit each time a defective item is found, and the value of β is increased by one unit each time a nondefective item is found.

Definition 7.3.1 Conjugate Family/Hyperparameters. Let $X_1, X_2, \ldots$ be conditionally i.i.d. given θ with common p.f. or p.d.f. $f(x|\theta)$. Let Ψ be a family of possible distributions over the parameter space Ω. Suppose that, no matter which prior distribution ξ we choose from Ψ, no matter how many observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ we observe, and no matter what are their observed values $\boldsymbol{x} = (x_1, \ldots, x_n)$, the posterior distribution $\xi(\theta|\boldsymbol{x})$ is a member of Ψ. Then Ψ is called a *conjugate family of prior distributions* for samples from the distributions $f(x|\theta)$. It is also said that the family Ψ is *closed under sampling* from the distributions $f(x|\theta)$. Finally, if the distributions in Ψ are parametrized by further parameters, then the associated parameters for the prior distribution are called the *prior hyperparameters* and the associated parameters of the posterior distribution are called the *posterior hyperparameters*.

Theorem 7.3.1 says that the family of beta distributions is a conjugate family of prior distributions for samples from a Bernoulli distribution. If the prior distribution of θ is a beta distribution, then the posterior distribution at each stage of sampling will also be a beta distribution, regardless of the observed values in the sample. Also, the family of beta distributions is closed under sampling from Bernoulli distributions. The parameters α and β in Theorem 7.3.1 are the prior hyperparameters. The corresponding parameters of the posterior distributions ($\alpha + \sum_{i=1}^n x_i$ and $\beta + n - \sum_{i=1}^n x_i$) are the posterior hyperparameters. The statistic $\sum_{i=1}^n X_i$ is needed to compute the posterior distribution, hence it will be needed to perform any inference based on the posterior distribution. Exercises 23 and 24 introduce a general collection of p.d.f.'s $f(x|\theta)$ for which conjugate families of priors exist. Most of the familiar named distributions are covered by these exercises. The various uniform distributions are notable exceptions.

Example 7.3.2 The Variance of the Posterior Beta Distribution. Suppose that the proportion θ of defective items in a large shipment is unknown, the prior distribution of θ is the uniform distribution on the interval [0, 1], and items are to be selected at random from the shipment and inspected until the variance of the posterior distribution of θ

has been reduced to the value 0.01 or less. We shall determine the total number of defective and nondefective items that must be obtained before the sampling process is stopped.

As stated in Sec. 5.8, the uniform distribution on the interval [0, 1] is the beta distribution with parameters 1 and 1. Therefore, after y defective items and z nondefective items have been obtained, the posterior distribution of θ will be the beta distribution with $\alpha = y + 1$ and $\beta = z + 1$. It was shown in Theorem 5.8.3 that the variance of the beta distribution with parameters α and β is $\alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]$. Therefore, the variance V of the posterior distribution of θ will be

$$V = \frac{(y+1)(z+1)}{(y+z+2)^2(y+z+3)}.$$

Sampling is to stop as soon as the number of defectives y and the number of nondefectives z that have been obtained are such that $V \le 0.01$. It can be shown (see Exercise 2) that it will not be necessary to select more than 22 items, but it is necessary to select at least seven items. ◀

Example 7.3.3

Glove Use by Nurses. Friedland et al. (1992) studied 23 nurses in an inner-city hospital before and after an educational program on the importance of wearing gloves. They recorded whether or not the nurses wore gloves during procedures in which they might come in contact with bodily fluids. Before the educational program the nurses were observed during 51 procedures, and they wore gloves in only 13 of them. Let θ be the probability that a nurse will wear gloves two months after the educational program. We might be interested in how θ compares to 13/51, the observed proportion before the program.

We shall consider two different prior distributions for θ in order to see how sensitive the posterior distribution of θ is to the choice of prior distribution. The first prior distribution will be uniform on the interval [0, 1], which is also the beta distribution with parameters 1 and 1. The second prior distribution will be the beta distribution with parameters 13 and 38. This second prior distribution has much smaller variance than the first and has its mean at 13/51. Someone holding the second prior distribution believes fairly strongly that the educational program will have no noticeable effect.

Two months after the educational program, 56 procedures were observed with the nurses wearing gloves in 50 of them. The posterior distribution of θ, based on the first prior, would then be the beta distribution with parameters $1 + 50 = 51$ and $1 + 6 = 7$. In particular, the posterior mean of θ is $51/(51 + 7) = 0.88$, and the posterior probability that $\theta > 2 \times 13/51$ is essentially 1. Based on the second prior, the posterior distribution would be the beta distribution with parameters $13 + 50 = 63$ and $38 + 6 = 44$. The posterior mean would be 0.59, and the posterior probability that $\theta > 2 \times 13/51$ is 0.95. So, even to someone who was initially skeptical, the educational program seems to have been quite effective. The probability is quite high that nurses are at least twice as likely to wear gloves after the program as they were before.

Figure 7.3 shows the p.d.f.'s of both of the posterior distributions computed above. The distributions are clearly very different. For example, the first posterior gives probability greater than 0.99 that $\theta > 0.7$, while the second gives probability less than 0.001 to $\theta > 0.7$. However, since we are only interested in the probability that $\theta > 2 \times 13/51 = 0.5098$, we see that both posteriors agree that this probability is quite large. ◀

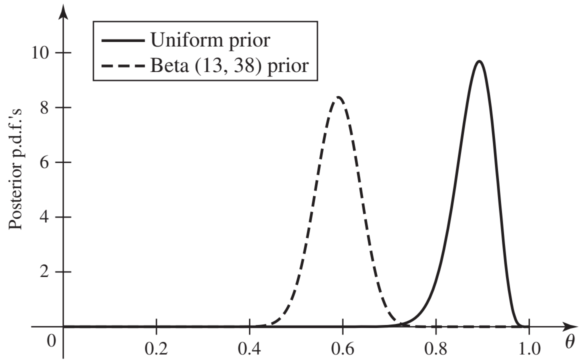


Figure 7.3 Posterior p.d.f.'s in Example 7.2.6. The curves are labeled by the prior that led to the corresponding posterior.

Sampling from a Poisson Distribution

Example 7.3.4 Customer Arrivals. A store owner models customer arrivals as a Poisson process with unknown rate θ per hour. She assigns θ a gamma prior distribution with parameters 3 and 2. Let X be the number of customers that arrive in a specific one-hour period. If $X = 3$ is observed, the store owner wants to update the distribution of θ. ◄

When samples are taken from a Poisson distribution, the family of gamma distributions is a conjugate family of prior distributions. This relationship is shown in the next theorem.

Theorem 7.3.2 Suppose that $X_1, \dots, X_n$ form a random sample from the Poisson distribution with mean $\theta > 0$, and θ is unknown. Suppose also that the prior distribution of θ is the gamma distribution with parameters $\alpha > 0$ and $\beta > 0$. Then the posterior distribution of θ, given that $X_i = x_i$ $(i = 1, \dots, n)$, is the gamma distribution with parameters $\alpha + \sum_{i=1}^{n} x_i$ and $\beta + n$.

Proof Let $y = \sum_{i=1}^{n} x_i$. Then the likelihood function $f_n(\boldsymbol{x}|\theta)$ satisfies the relation

$$f_n(\boldsymbol{x}|\theta) \propto e^{-n\theta}\theta^{y}.$$

In this relation, a factor that involves $\boldsymbol{x}$ but does not depend on θ has been dropped from the right side. Furthermore, the prior p.d.f. of θ has the form

$$\xi(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta} \quad \text{for } \theta > 0.$$

Since the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ is proportional to $f_n(\boldsymbol{x}|\theta)\xi(\theta)$, it follows that

$$\xi(\theta|\boldsymbol{x}) \propto \theta^{\alpha+y-1}e^{-(\beta+n)\theta} \quad \text{for } \theta > 0.$$

The right side of this relation can be recognized as being, except for a constant factor, the p.d.f. of the gamma distribution with parameters $\alpha + y$ and $\beta + n$. Therefore, the posterior distribution of θ is as specified in the theorem. ■

In Theorem 7.3.2, the numbers α and β are the prior hyperparameters, while $\alpha + \sum_{i=1}^{n} x_i$ and $\beta + n$ are the posterior hyperparameters. Note that the statistic $Y = \sum_{i=1}^{n} X_i$ is used to compute the posterior distribution of θ, and hence it will be part of any inference based on the posterior.

Example 7.3.5 Customer Arrivals. In Example 7.3.4, we can apply Theorem 7.3.2 with $n = 1$, $\alpha = 3$, $\beta = 2$, and $x_1 = 3$. The posterior distribution of θ given $X = 3$ is the gamma distribution with parameters 6 and 3. ◀

Example 7.3.6 The Variance of the Posterior Gamma Distribution. Consider a Poisson distribution for which the mean θ is unknown, and suppose that the prior p.d.f. of θ is as follows:

$$\xi(\theta) = \begin{cases} 2e^{-2\theta} & \text{for } \theta > 0, \\ 0 & \text{for } \theta \le 0. \end{cases}$$

Suppose also that observations are to be taken at random from the given Poisson distribution until the variance of the posterior distribution of θ has been reduced to the value 0.01 or less. We shall determine the number of observations that must be taken before the sampling process is stopped.

The given prior p.d.f. $\xi(\theta)$ is the p.d.f. of the gamma distribution with prior hyperparameters $\alpha = 1$ and $\beta = 2$. Therefore, after we have obtained n observed values $x_1, \dots, x_n$, the sum of which is $y = \sum_{i=1}^n x_i$, the posterior distribution of θ will be the gamma distribution with posterior hyperparameters $y + 1$ and $n + 2$. It was shown in Theorem 5.4.2 that the variance of the gamma distribution with parameters α and β is α/β^2. Therefore, the variance V of the posterior distribution of θ will be

$$V = \frac{y+1}{(n+2)^2}.$$

Sampling is to stop as soon as the sequence of observed values $x_1, \dots, x_n$ is such that $V \le 0.01$. Unlike Example 7.3.2, there is no uniform bound on how large n needs to be because y can be arbitrarily large no matter what n is. Clearly, it takes at least $n = 8$ observations before $V \le 0.01$. ◀

Sampling from a Normal Distribution

Example 7.3.7 Automobile Emissions. Consider again the sampling of automobile emissions, in particular oxides of nitrogen, described in Example 5.6.1 on page 302. Prior to observing the data, suppose that an engineer believed that each emissions measurement had the normal distribution with mean θ and standard deviation 0.5 but that θ was unknown. The engineer's uncertainty about θ might be described by another normal distribution with mean 2.0 and standard deviation 1.0. After seeing the data in Fig. 5.1, how would this engineer describe her uncertainty about θ? ◀

When samples are taken from a normal distribution for which the value of the mean θ is unknown but the value of the variance σ^2 is known, the family of normal distributions is itself a conjugate family of prior distributions, as is shown in the next theorem.

Theorem 7.3.3 Suppose that $X_1, \dots, X_n$ form a random sample from a normal distribution for which the value of the mean θ is unknown and the value of the variance $\sigma^2 > 0$ is known. Suppose also that the prior distribution of θ is the normal distribution with mean μ_0 and variance v_0^2. Then the posterior distribution of θ given that $X_i = x_i$ $(i = 1, \dots, n)$ is the normal distribution with mean μ_1 and variance v_1^2 where

$$\mu_1 = \frac{\sigma^2\mu_0 + n v_0^2 \overline{x}_n}{\sigma^2 + n v_0^2} \tag{7.3.1}$$

and

$$v_1^2 = \frac{\sigma^2 v_0^2}{\sigma^2 + n v_0^2}. \tag{7.3.2}$$

Proof The likelihood function. $f_n(\boldsymbol{x}|\theta)$ has the form

$$f_n(\boldsymbol{x}|\theta) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right].$$

Here a constant factor has been dropped from the right side. The method of completing the square (see Exercise 24 in Sec. 5.6) tells us that

$$\sum_{i=1}^{n}(x_i-\theta)^2 = n(\theta-\overline{x}_n)^2 + \sum_{i=1}^{n}(x_i-\overline{x}_n)^2.$$

By omitting a factor that involves $x_1, \ldots, x_n$ but does not depend on θ, we may rewrite $f_n(\boldsymbol{x}|\theta)$ in the following form:

$$f_n(\boldsymbol{x}|\theta) \propto \exp\left[-\frac{n}{2\sigma^2}(\theta-\overline{x}_n)^2\right].$$

Since the prior p.d.f. $\xi(\theta)$ has the form

$$\xi(\theta) \propto \exp\left[-\frac{1}{2v_0^2}(\theta-\mu_0)^2\right],$$

it follows that the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ satisfies the relation

$$\xi(\theta|\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2}\left[\frac{n}{\sigma^2}(\theta-\overline{x}_n)^2 + \frac{1}{v_0^2}(\theta-\mu_0)^2\right]\right\}.$$

If μ_1 and v_1^2 are as specified in Eqs. (7.3.1) and (7.3.2), completing the square again establishes the following identity:

$$\frac{n}{\sigma^2}(\theta-\overline{x}_n)^2 + \frac{1}{v_0^2}(\theta-\mu_0)^2 = \frac{1}{v_1^2}(\theta-\mu_1)^2 + \frac{n}{\sigma^2+nv_0^2}(\overline{x}_n-\mu_0)^2.$$

Since the final term on the right side of this equation does not involve θ, it can be absorbed in the proportionality factor, and we obtain the relation

$$\xi(\theta|\boldsymbol{x}) \propto \exp\left[-\frac{1}{2v_1^2}(\theta-\mu_1)^2\right].$$

The right side of this relation can be recognized as being, except for a constant factor, the p.d.f. of the normal distribution with mean μ_1 and variance v_1^2. Therefore, the posterior distribution of θ is as specified in the theorem. ■

In Theorem 7.3.3, the numbers μ_0 and v_0^2 are the prior hyperparameters, while μ_1 and v_1^2 are the posterior hyperparameters. Notice that the statistic $\overline{X}_n$ is used in the construction of the posterior distribution, and hence will play a role in any inference based on the posterior.

Example 7.3.8 Automobile Emissions. We can apply Theorem 7.3.3 to answer the question at the end of Example 7.3.7. In the notation of the theorem, we have $n = 46$, $\sigma^2 = 0.5^2 = 0.25$,

$\mu_0 = 2$, and $v^2 = 1.0$. The average of the 46 measurements is $\overline{x}_n = 1.329$. The posterior distribution of θ is then the normal distribution with mean and variance given by

$$\mu_1 = \frac{0.25 \times 2 + 46 \times 1 \times 1.329}{0.25 + 46 \times 1} = 1.333,$$

$$v_1^2 = \frac{0.25 \times 1}{0.25 + 46 \times 1} = 0.0054. \quad ◀$$

The mean μ_1 of the posterior distribution of θ, as given in Eq. (7.3.1), can be rewritten as follows:

$$\mu_1 = \frac{\sigma^2}{\sigma^2 + nv_0^2}\mu_0 + \frac{nv_0^2}{\sigma^2 + nv_0^2}\overline{x}_n. \tag{7.3.3}$$

It can be seen from Eq. (7.3.3) that μ_1 is a weighted average of the mean μ_0 of the prior distribution and the sample mean $\overline{x}_n$. Furthermore, it can be seen that the relative weight given to $\overline{x}_n$ satisfies the following three properties: (1) For fixed values of v_0^2 and σ^2, the larger the sample size n, the greater will be the relative weight that is given to $\overline{x}_n$. (2) For fixed values of v_0^2 and n, the larger the variance σ^2 of each observation in the sample, the smaller will be the relative weight that is given to $\overline{x}_n$. (3) For fixed values of σ^2 and n, the larger the variance v_0^2 of the prior distribution, the larger will be the relative weight that is given to $\overline{x}_n$.

Moreover, it can be seen from Eq. (7.3.2) that the variance v_1^2 of the posterior distribution of θ depends on the number n of observations that have been taken but does not depend on the magnitudes of the observed values. Suppose, therefore, that a random sample of n observations is to be taken from a normal distribution for which the value of the mean θ is unknown, the value of the variance is known, and the prior distribution of θ is a specified normal distribution. Then, before any observations have been taken, we can use Eq. (7.3.2) to calculate the actual value of the variance v_1^2 of the posterior distribution. However, the value of the mean μ_1 of the posterior distribution will depend on the observed values that are obtained in the sample. The fact that the variance of the posterior distribution depends only on the number of observations is due to the assumption that the variance σ^2 of the individual observations is known. In Sec. 8.6, we shall relax this assumption.

Example 7.3.9 The Variance of the Posterior Normal Distribution. Suppose that observations are to be taken at random from the normal distribution with mean θ and variance 1, and that θ is unknown. Assume that the prior distribution of θ is a normal distribution with variance 4. Also, observations are to be taken until the variance of the posterior distribution of θ has been reduced to the value 0.01 or less. We shall determine the number of observations that must be taken before the sampling process is stopped.

It follows from Eq. (7.3.2) that after n observations have been taken, the variance v_1^2 of the posterior distribution of θ will be

$$v_1^2 = \frac{4}{4n+1}.$$

Therefore, the relation $v_1^2 \leq 0.01$ will be satisfied if and only if $n \geq 99.75$. Hence, the relation $v_1^2 \leq 0.01$ will be satisfied after 100 observations have been taken and not before then. ◀

Example 7.3.10 Calorie Counts on Food Labels. Allison, Heshka, Sepulveda, and Heymsfield (1993) sampled 20 nationally prepared foods and compared the stated calorie contents per

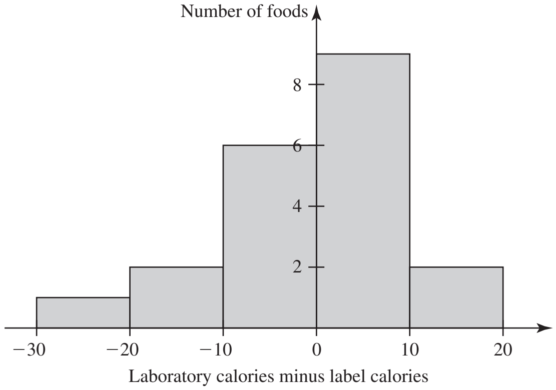


Figure 7.4 Histogram of percentage differences between observed and advertised calories in Example 7.3.10.

gram from the labels to calorie contents determined in the laboratory. Figure 7.4 is a histogram of the percentage differences between the observed laboratory calorie measurements and the advertised calorie contents on the labels of the foods. Suppose that we model the conditional distribution of the differences given θ as the normal distribution with mean θ and variance 100. (In this section, we assume that the variance is known. In Sec. 8.6, we will be able to deal with the case in which the mean and the variance are treated as random variables with a joint distribution.) We will use a prior distribution for θ that is the normal distribution with mean 0 and a variance of 60. The data $\boldsymbol{X}$ comprise the collection of 20 differences in Fig. 7.4, whose average is 0.125. The posterior distribution of θ would then be the normal distribution with mean

$$\mu_1 = \frac{100 \times 0 + 20 \times 60 \times 0.125}{100 + 20 \times 60} = 0.1154,$$

and variance

$$v_1^2 = \frac{100 \times 60}{100 + 20 \times 60} = 4.62.$$

For example, we might be interested in whether or not the packagers are systematically understating the calories in their food by at least 1 percent. This would correspond to $\theta > 1$. Using Theorem 5.6.6, we can find

$$\Pr(\theta > 1|\boldsymbol{x}) = 1 - \Phi\left(\frac{1 - 0.1154)}{\sqrt{4.62}}\right) = 1 - \Phi(1.12) = 0.3403.$$

There is a nonnegligible, but not overwhelming, chance that the packagers are shaving a percent or more off of their labels. ◀

Sampling from an Exponential Distribution

Example 7.3.11 **Lifetimes of Electronic Components.** In Example 7.2.1, suppose that we observe the lifetimes of three components, $X_1 = 3$, $X_2 = 1.5$, and $X_3 = 2.1$. These were modeled as i.i.d. exponential random variables given θ. Our prior distribution for θ was the gamma distribution with parameters 1 and 2. What is the posterior distribution of θ given these observed lifetimes? ◀

When sampling from an exponential distribution for which the value of the parameter θ is unknown, the family of gamma distributions serves as a conjugate family of prior distributions, as shown in the next theorem.

Theorem 7.3.4 Suppose that $X_1, \dots, X_n$ form a random sample from the exponential distribution with parameter $\theta > 0$ that is unknown. Suppose also that the prior distribution of θ is the gamma distribution with parameters $\alpha > 0$ and $\beta > 0$. Then the posterior distribution of θ given that $X_i = x_i$ $(i = 1, \dots, n)$ is the gamma distribution with parameters $\alpha + n$ and $\beta + \sum_{i=1}^{n} x_i$.

Proof Again, let $y = \sum_{i=1}^{n} x_i$. Then the likelihood function $f_n(\boldsymbol{x}|\theta)$ is

$$f_n(\boldsymbol{x}|\theta) = \theta^n e^{-\theta y}.$$

Also, the prior p.d.f. $\xi(\theta)$ has the form

$$\xi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta} \quad \text{for } \theta > 0.$$

It follows, therefore, that the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ has the form

$$\xi(\theta|\boldsymbol{x}) \propto \theta^{\alpha+n-1} e^{-(\beta+y)\theta} \quad \text{for } \theta > 0.$$

The right side of this relation can be recognized as being, except for a constant factor, the p.d.f. of the gamma distribution with parameters $\alpha + n$ and $\beta + y$. Therefore, the posterior distribution of θ is as specified in the theorem. ■

The posterior distribution of θ in Theorem 7.3.4 depends on the observed value of the statistic $Y = \sum_{i=1}^{n} X_i$; hence, every inference about θ based on the posterior distribution will depend on the observed value of Y.

Example 7.3.12 Lifetimes of Electronic Components. In Example 7.3.11, we can apply Theorem 7.3.4 to find the posterior distribution. In the notation of the theorem and its proof, we have $n = 3$, $\alpha = 1$, $\beta = 2$, and

$$y = \sum_{i=1}^{n} x_i = 3 + 1.5 + 2.1 = 6.6.$$

The posterior distribution of θ is then the gamma distribution with parameters $\alpha = 1 + 3 = 4$ and $\beta = 2 + 6.6 = 8.6$. ◄

The reader should note that Theorem 7.3.4 would have greatly shortened the derivation of the posterior distribution in Example 7.2.6.

Improper Prior Distributions

In Sec. 7.2, we mentioned improper priors as expedients that try to capture the idea that there is much more information in the data than is captured in our prior distribution. Each of the conjugate families that we have seen in this section has an improper prior as a limiting case.

Example 7.3.13 A Clinical Trial. What we illustrate here will apply to all examples in which the data comprise a conditionally i.i.d. sample (given θ) from the Bernoulli distribution with parameter θ. Consider the subjects in the imipramine group in Example 2.1.4. The proportion of successes among all patients who might get imipramine had been called P in earlier examples, but let us call it θ this time in keeping with the general notation

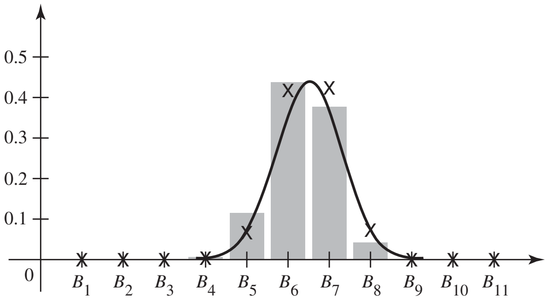


Figure 7.5 The posterior probabilities from Examples 2.3.7 (X) and 2.3.8 (bars) together with the posterior p.d.f. from Example 7.3.13 (solid line).

of this chapter. Suppose that θ has the beta distribution with parameters α and β, a general conjugate prior. There are $n = 40$ patients in the imipramine group, and 22 of them are successes. The posterior distribution of θ is the beta distribution with parameters $\alpha + 22$ and $\beta + 18$, as we saw in Theorem 7.3.1. The mean of the posterior distribution is $(\alpha + 22)/(\alpha + \beta + 40)$. If α and β are small, then the posterior mean is close to 22/40, which is the observed proportion of successes. Indeed, if $\alpha = \beta = 0$, which does not correspond to a real beta distribution, then the posterior mean is exactly 22/40. However, we can look at what happens as α and β get close to 0. The beta p.d.f. (ignoring the constant factor) is $\theta^{\alpha-1}(1-\theta)^{\beta-1}$. We can set $\alpha = \beta = 0$ and pretend that $\xi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$ is the prior p.d.f. of θ. The likelihood function is $f_{40}(\boldsymbol{x}|\theta) = \binom{40}{22}\theta^{22}(1-\theta)^{18}$. We can ignore the constant factor $\binom{40}{22}$ and obtain the product

$$\xi(\theta|\boldsymbol{x}) \propto \theta^{21}(1-\theta)^{17}, \quad \text{for } 0 < \theta < 1.$$

This is easily recognized as being the same as the p.d.f. of the beta distribution with parameters 22 and 18 except for a constant factor. So, if we use the improper "beta distribution" prior with prior hyperparameters 0 and 0, we get the beta posterior distribution for θ with posterior hyperparameters 22 and 18. Notice that Theorem 7.3.1 yields the correct posterior distribution even in this improper prior case. Figure 7.5 adds the p.d.f. of the posterior beta distribution calculated here to Fig. 2.4 which depicted the posterior probabilities for two different discrete prior distributions. All three posteriors are pretty close. ◄

Definition 7.3.2 Improper Prior. Let ξ be a nonnegative function whose domain includes the parameter space of a statistical model. Suppose that $\int \xi(\theta)d\theta = \infty$. If we pretend as if $\xi(\theta)$ is the prior p.d.f. of θ, then we are using an *improper prior* for θ.

Definition 7.3.2 is not of much use in determining an improper prior to use in a particular application. There are many methods for choosing an improper prior, and the hope is that they all lead to similar posterior distributions so that it does not much matter which of them one chooses. The most straightforward method for choosing an improper prior is to start with the family of conjugate prior distributions, if there is such a family. In most cases, if the parameterization of the conjugate family (prior hyperparameters) is chosen carefully, the posterior hyperparameters will each equal the corresponding prior hyperparameter plus a statistic. One would then replace each of those prior hyperparameters by 0 in the formula for the prior p.d.f. This generally results in a function that satisfies Definition 7.3.2. In Example 7.3.13, each of the posterior hyperparameters were equal to the corresponding prior hyperparameters plus some statistic. In that example, we replaced both prior hyperparameters by 0 to obtain the improper prior. Here are some more examples. The method just

described needs to be modified if one chooses an "inconvenient" parameterization of the conjugate prior, as in Example 7.3.15 below.

Example 7.3.14

Prussian Army Deaths. Bortkiewicz (1898) counted the numbers of Prussian soldiers killed by horsekick (a more serious problem in the nineteenth century than it is today) in 14 army units for each of 20 years, a total of 280 counts. The 280 counts have the following values: 144 counts are 0, 91 counts are 1, 32 counts are 2, 11 counts are 3, and 2 counts are 4. No unit suffered more than four deaths by horsekick during any one year. (These data were reported and analyzed by Winsor, 1947.) Suppose that we were going to model the 280 counts as a random sample of Poisson random variables $X_1, \dots, X_{280}$ with mean θ conditional on the parameter θ. A conjugate prior would be a member of the gamma family with prior hyperparameters α and β. Theorem 7.3.2 says that the posterior distribution of θ would be the gamma distribution with posterior hyperparameters $\alpha + 196$ and $\beta + 280$, since the sum of the 280 counts equals 196. Unless either α or β is very large, the posterior gamma distribution is nearly the same as the gamma distribution with posterior hyperparameters 196 and 280. This posterior distribution would seem to be the result of using a conjugate prior with prior hyperparameters 0 and 0. Ignoring the constant factor, the p.d.f. of the gamma distribution with parameters α and β is $\theta^{\alpha-1}e^{\beta\theta}$ for $\theta > 0$. If we let $\alpha = 0$ and $\beta = 0$ in this formula, we get the improper prior "p.d.f." $\xi(\theta) = \theta^{-1}$ for $\theta > 0$. Pretending as if this really were a prior p.d.f. and applying Bayes' theorem for random variables (Theorem 3.6.4) would yield

$$\xi(\theta|\boldsymbol{x}) \propto \theta^{195}e^{-280\theta}, \quad \text{for } \theta > 0.$$

This is easily recognized as being the p.d.f. of the gamma distribution with parameters 196 and 280, except for a constant factor. The result in this example applies to all cases in which we model data with Poisson distributions. The improper "gamma distribution" with prior hyperparameters 0 and 0 can be used in Theorem 7.3.2, and the conclusion will still hold. ◀

Example 7.3.15

Failure Times of Ball Bearings. Suppose that we model the 23 logarithms of failure times of ball bearings from Example 5.6.9 as normal random variables $X_1, \dots, X_{23}$ with mean θ and variance 0.25. A conjugate prior for θ would be the normal distribution with mean μ_0 and variance v_0^2 for some μ_0 and v_0^2. The average of the 23 log-failure times is 4.15, so the posterior distribution of θ would be the normal distribution with mean $\mu_1 = (0.25\mu_0 + 23 \times 4.15v_0^2)/(0.25 + 23v_0^2)$ and variance $v_1^2 = (0.25v_0^2)/(0.25 + 23v_0^2)$. If we let $v_0^2 \to \infty$ in the formulas for μ_1 and v_1^2, we get $\mu_1 \to 4.15$ and $v_1^2 \to 0.25/23$. Having infinite variance for the prior distribution of θ is like saying that θ is equally likely to be anywhere on the real number line. This same thing happens in every example in which we model data $X_1, \dots, X_n$ as a random sample from the normal distribution with mean θ and known variance σ^2 conditional on θ. If we use an improper "normal distribution" prior with variance ∞ (the prior mean does not matter), the calculation in Theorem 7.3.3 would yield a posterior distribution that is the normal distribution with mean $\overline{x}_n$ and variance σ^2/n. The improper prior "p.d.f." in this case is $\xi(\theta)$ equal to a constant.

This example would be an application of the method described after Definition 7.3.2 if we had described the conjugate prior distribution in terms of the following "more convenient" hyperparameters: 1 over the variance $u_0 = 1/v_0^2$ and the mean over the variance $t_0 = \mu_0/v_0^2$. In terms of these hyperparameters, the posterior distribution has 1 over its variance equal to $u_1 = u_0 + n/0.25$ and mean over variance equal to $t_1 = \mu_1/v_1^2 = t_0 + 23 \times 4.15/0.25$. Each of u_1 and t_1 has the form of the cor-

responding prior hyperparamter plus a statistic. The improper prior with $u_0 = t_0 = 0$ also has $\xi(\theta)$ equal to a constant. ◀

There are improper priors for other sampling models, also. The reader can verify (in Exercise 21) that the "gamma distribution" with parameters 0 and 0 leads to results similar to those in Example 7.3.14 when the data are a random sample from an exponential distribution. Exercises 23 and 24 introduce a general collection of p.d.f.'s $f(x|\theta)$ for which it is easy to construct improper priors.

Improper priors were introduced for cases in which the observed data contain much more information than is represented by our prior distribution. Implicitly, we are assuming that the data are rather informative. When the data do not contain much information, improper priors may be higly inappropriate.

Example 7.3.16

Very Rare Events. In Example 5.4.7, we discussed a drinking water contaminant known as cryptosporidium that generally occurs in very low concentrations. Suppose that a water authority models the oocysts of cryptosporidium in the water supply as a Poisson process with rate of θ oocysts per liter. They decide to sample 25 liters of water to learn about θ. Suppose that they use the improper gamma prior with "p.d.f." θ^{-1}. (This is the same improper prior used in Example 7.3.14.) If the 25-liter sample contains no oocysts, the water authority would be led to a posterior distribution for θ that was the gamma distribution with parameters 0 and 5, which is not a real distribution. No matter how many liters are sampled, the posterior distribution will not be a real distribution until at least one oocyst is observed. When sampling for rare events, one might be forced to quantify prior information in the form a proper prior distribution in order to be able to make inferences based on the posterior distribution. ◀

Summary

For each of several different statistical models for data given the parameter, we found a conjugate family of distributions for the parameter. These families have the property that if the prior distribution is chosen from the family, then the posterior distribution is a member of the family. For data with distributions related to the Bernoulli, such as binomial, geometric, and negative binomial, the conjugate family for the success probability parameter is the family of beta distributions. For data with distributions related to the Poisson process, such as Poisson, gamma (with known first parameter), and exponential, the conjugate family for the rate parameter is the family of gamma distributions. For data having a normal distribution with known variance, the conjugate family for the mean is the normal family. We also described the use of improper priors. Improper priors are not true probability distributions, but if we pretend that they are, we will compute posterior distributions that approximate the posteriors that we would have obtained using proper conjugate priors with extreme values of the prior hyperparameters.

Exercises

1. Consider again the situation described in Example 7.3.10. Once again, suppose that the prior distribution of θ is a normal distribution with mean 0, but this time let the prior variance be $v^2 > 0$. If the posterior mean of θ is 0.12, what value of v^2 was used?

2. Show that in Example 7.3.2 it must be true that $V \le 0.01$ after 22 items have been selected. Also show that $V > 0.01$ until at least seven items have been selected.

3. Suppose that the proportion θ of defective items in a large shipment is unknown and that the prior distribution

of θ is the beta distribution with parameters 2 and 200. If 100 items are selected at random from the shipment and if three of these items are found to be defective, what is the posterior distribution of θ?

4. Consider again the conditions of Exercise 3. Suppose that after a certain statistician has observed that there were three defective items among the 100 items selected at random, the posterior distribution that she assigns to θ is a beta distribution for which the mean is 2/51 and the variance is $98/[(51)^2(103)]$. What prior distribution had the statistician assigned to θ?

5. Suppose that the number of defects in a 1200-foot roll of magnetic recording tape has a Poisson distribution for which the value of the mean θ is unknown and that the prior distribution of θ is the gamma distribution with parameters $\alpha = 3$ and $\beta = 1$. When five rolls of this tape are selected at random and inspected, the numbers of defects found on the rolls are 2, 2, 6, 0, and 3. Determine the posterior distribution of θ.

6. Let θ denote the average number of defects per 100 feet of a certain type of magnetic tape. Suppose that the value of θ is unknown and that the prior distribution of θ is the gamma distribution with parameters $\alpha = 2$ and $\beta = 10$. When a 1200-foot roll of this tape is inspected, exactly four defects are found. Determine the posterior distribution of θ.

7. Suppose that the heights of the individuals in a certain population have a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2 inches. Suppose also that the prior distribution of θ is a normal distribution for which the mean is 68 inches and the standard deviation is 1 inch. If 10 people are selected at random from the population, and their average height is found to be 69.5 inches, what is the posterior distribution of θ?

8. Consider again the problem described in Exercise 7.

- **a.** Which interval 1-inch long had the highest prior probability of containing the value of θ?
- **b.** Which interval 1-inch long has the highest posterior probability of containing the value of θ?
- **c.** Find the values of the probabilities in parts (a) and (b).

9. Suppose that a random sample of 20 observations is taken from a normal distribution for which the value of the mean θ is unknown and the variance is 1. After the sample values have been observed, it is found that $\overline{X}_n = 10$, and that the posterior distribution of θ is a normal distribution for which the mean is 8 and the variance is 1/25. What was the prior distribution of θ?

10. Suppose that a random sample is to be taken from a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2, and the prior distribution of θ is a normal distribution for which the standard deviation is 1. What is the smallest number of observations that must be included in the sample in order to reduce the standard deviation of the posterior distribution of θ to the value 0.1?

11. Suppose that a random sample of 100 observations is to be taken from a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2, and the prior distribution of θ is a normal distribution. Show that no matter how large the standard deviation of the prior distribution is, the standard deviation of the posterior distribution will be less than 1/5.

12. Suppose that the time in minutes required to serve a customer at a certain facility has an exponential distribution for which the value of the parameter θ is unknown and that the prior distribution of θ is a gamma distribution for which the mean is 0.2 and the standard deviation is 1. If the average time required to serve a random sample of 20 customers is observed to be 3.8 minutes, what is the posterior distribution of θ?

13. For a distribution with mean $\mu \neq 0$ and standard deviation $\sigma > 0$, the *coefficient of variation* of the distribution is defined as $\sigma/|\mu|$. Consider again the problem described in Exercise 12, and suppose that the coefficient of variation of the prior gamma distribution of θ is 2. What is the smallest number of customers that must be observed in order to reduce the coefficient of variation of the posterior distribution to 0.1?

14. Show that the family of beta distributions is a conjugate family of prior distributions for samples from a negative binomial distribution with a known value of the parameter r and an unknown value of the parameter p $(0 < p < 1)$.

15. Let $\xi(\theta)$ be a p.d.f. that is defined as follows for constants $\alpha > 0$ and $\beta > 0$:

$$\xi(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta} & \text{for } \theta > 0, \\ 0 & \text{for } \theta \le 0. \end{cases}$$

A distribution with this p.d.f. is called an *inverse gamma distribution*.

- **a.** Verify that $\xi(\theta)$ is actually a p.d.f. by verifying that $\int_0^\infty \xi(\theta)\, d\theta = 1$.
- **b.** Consider the family of probability distributions that can be represented by a p.d.f. $\xi(\theta)$ having the given form for all possible pairs of constants $\alpha > 0$ and $\beta > 0$. Show that this family is a conjugate family of prior distributions for samples from a normal distribution with a known value of the mean μ and an unknown value of the variance θ.

16. Suppose that in Exercise 15 the parameter is taken as the standard deviation of the normal distribution, rather than the variance. Determine a conjugate family of prior distributions for samples from a normal distribution with

a known value of the mean μ and an unknown value of the standard deviation σ.

17. Suppose that the number of minutes a person must wait for a bus each morning has the uniform distribution on the interval $[0, \theta]$, where the value of the endpoint θ is unknown. Suppose also that the prior p.d.f. of θ is as follows:

$$\xi(\theta) = \begin{cases} \frac{192}{\theta^4} & \text{for } \theta \ge 4, \\ 0 & \text{otherwise.} \end{cases}$$

If the observed waiting times on three successive mornings are 5, 3, and 8 minutes, what is the posterior p.d.f. of θ?

18. The Pareto distribution with parameters x_0 and α ($x_0 > 0$ and $\alpha > 0$) is defined in Exercise 16 of Sec. 5.7. Show that the family of Pareto distributions is a conjugate family of prior distributions for samples from a uniform distribution on the interval $[0, \theta]$, where the value of the endpoint θ is unknown.

19. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. $f(x|\theta)$ is as follows:

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose also that the value of the parameter θ is unknown ($\theta > 0$), and the prior distribution of θ is the gamma distribution with parameters α and β ($\alpha > 0$ and $\beta > 0$). Determine the mean and the variance of the posterior distribution of θ.

20. Suppose that we model the lifetimes (in months) of electronic components as independent exponential random variables with unknown parameter β. We model β as having the gamma distribution with parameters a and b. We believe that the mean lifetime is four months before we see any data. If we were to observe 10 components with an average observed lifetime of six months, we would then claim that the mean lifetime is five months. Determine a and b. *Hint:* Use Exercise 21 in Sec. 5.7.

21. Suppose that $X_1, \ldots, X_n$ form a random sample from the exponential distribution with parameter θ. Let the prior distribution of θ be improper with "p.d.f." $1/\theta$ for $\theta > 0$. Find the posterior distribution of θ and show that the posterior mean of θ is $1/\overline{x}_n$.

22. Consider the data in Example 7.3.10. This time, suppose that we use the improper prior "p.d.f." $\xi(\theta) = 1$ (for all θ). Find the posterior distribution of θ and the posterior probability that $\theta > 1$.

23. Consider a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where θ belongs to some parameter space Ω. It is said that the family of distributions obtained by letting θ vary over all values in Ω is an *exponential family*, or a *Koopman-Darmois family*, if $f(x|\theta)$ can be written as follows for $\theta \in \Omega$ and all values of x:

$$f(x|\theta) = a(\theta)b(x)\exp[c(\theta)\,d(x)].$$

Here $a(\theta)$ and $c(\theta)$ are arbitrary functions of θ, and $b(x)$ and $d(x)$ are arbitrary functions of x. Let

$$H = \left\{(\alpha, \beta) : \int_{\Omega} a(\theta)^{\alpha} \exp[c(\theta)\,\beta] d\theta < \infty\right\}.$$

For each $(\alpha, \beta) \in H$, let

$$\xi_{\alpha,\beta}(\theta) = \frac{a(\theta)^{\alpha}\exp[c(\theta)\,\beta]}{\int_{\Omega} a(\eta)^{\alpha}\exp[c(\eta)\,\beta] d\eta},$$

and let Ψ be the set of all probability distributions that have p.d.f.'s of the form $\xi_{\alpha,\beta}(\theta)$ for some $(\alpha, \beta) \in H$.

- **a.** Show that Ψ is a conjugate family of prior distributions for samples from $f(x|\theta)$.
- **b.** Suppose that we observe a random sample of size n from the distribution with p.d.f. $f(x|\theta)$. If the prior p.d.f. of θ is ξ_{α_0,β_0}, show that the posterior hyperparameters are

$$\alpha_1 = \alpha_0 + n, \;\; \beta_1 = \beta_0 + \sum_{i=1}^{n} d(x_i).$$

24. Show that each of the following families of distributions is an exponential family, as defined in Exercise 23:

- **a.** The family of Bernoulli distributions with an unknown value of the parameter p
- **b.** The family of Poisson distributions with an unknown mean
- **c.** The family of negative binomial distributions for which the value of r is known and the value of p is unknown
- **d.** The family of normal distributions with an unknown mean and a known variance
- **e.** The family of normal distributions with an unknown variance and a known mean
- **f.** The family of gamma distributions for which the value of α is unknown and the value of β is known
- **g.** The family of gamma distributions for which the value of α is known and the value of β is unknown
- **h.** The family of beta distributions for which the value of α is unknown and the value of β is known
- **i.** The family of beta distributions for which the value of α is known and the value of β is unknown

25. Show that the family of uniform distributions on the intervals $[0, \theta]$ for $\theta > 0$ is *not* an exponential family as defined in Exercise 23. *Hint:* Look at the support of each uniform distribution.

26. Show that the family of discrete uniform distributions on the sets of integers $\{0, 1, \ldots, \theta\}$ for θ a nonnegative integer is *not* an exponential family as defined in Exercise 23.

7.4 Bayes Estimators

An estimator of a parameter is some function of the data that we hope is close to the parameter. A Bayes estimator is an estimator that is chosen to minimize the posterior mean of some measure of how far the estimator is from the parameter, such as squared error or absolute error.

Nature of an Estimation Problem

Example 7.4.1 Calorie Counts on Food Labels. In Example 7.3.10, we found the posterior distribution of θ, the mean percentage difference between measured and advertised calorie counts. A consumer group might wish to report a single number as an estimate of θ without specifying the entire distribution for θ. How to choose such a single-number estimate in general is the subject of this section. ◀

We begin with a definition that is appropriate for a real-valued parameter such as in Example 7.4.1. A more general definition will follow after we become more familiar with the concept of estimation.

Definition 7.4.1 Estimator/Estimate. Let $X_1, \dots, X_n$ be observable data whose joint distribution is indexed by a parameter θ taking values in a subset Ω of the real line. An *estimator* of the parameter θ is a real-valued function $\delta(X_1, \dots, X_n)$. If $X_1 = x_1, \dots, X_n = x_n$ are observed, then $\delta(x_1, \dots, x_n)$ is called the *estimate* of θ.

Notice that every estimator is, by nature of being a function of data, a statistic in the sense of Definition 7.1.4.

Because the value of θ must belong to the set Ω, it might seem reasonable to require that every possible value of an estimator $\delta(X_1, \dots, X_n)$ must also belong to Ω. We shall not require this restriction, however. If an estimator can take values outside of the parameter space Ω, the experimenter will need to decide in the specific problem whether that seems appropriate or not. It may turn out that every estimator that takes values only inside Ω has other even less desirable properties.

In Definition 7.4.1, we distinguished between the terms *estimator* and *estimate*. Because an estimator $\delta(X_1, \dots, X_n)$ is a function of the random variables $X_1, \dots, X_n$, the estimator itself is a random variable, and its probability distribution can be derived from the joint distribution of $X_1, \dots, X_n$, if desired. On the other hand, an *estimate* is a specific value $\delta(x_1, \dots, x_n)$ of the estimator that is determined by using specific observed values $x_1, \dots, x_n$. If we use the vector notation $\boldsymbol{X} = (X_1, \dots, X_n)$ and $\boldsymbol{x} = (x_1, \dots, x_n)$, then an estimator is a function $\delta(\boldsymbol{X})$ of the random vector $\boldsymbol{X}$, and an estimate is a specific value $\delta(\boldsymbol{x})$. It will often be convenient to denote an estimator $\delta(\boldsymbol{X})$ simply by the symbol δ.

Loss Functions

Example 7.4.2 Calorie Counts on Food Labels. In Example 7.4.1, the consumer group may feel that the farther their estimate $\delta(\boldsymbol{x})$ is from the true mean difference θ, the more embarassment and possible legal action they will encounter. Ideally, they would like to quantify the amount of negative repercussions as a function of θ and the estimate $\delta(\boldsymbol{x})$. Then they could have some idea how likely it is that they will encounter various levels of hassle as a result of their estimation. ◀

The foremost requirement of a good estimator δ is that it yield an estimate of θ that is close to the actual value of θ. In other words, a good estimator is one for which it is highly probable that the error $\delta(\boldsymbol{X}) - \theta$ will be close to 0. We shall assume that for each possible value of $\theta \in \Omega$ and each possible estimate a, there is a number $L(\theta, a)$ that measures the loss or cost to the statistician when the true value of the parameter is θ and her estimate is a. Typically, the greater the distance between a and θ, the larger will be the value of $L(\theta, a)$.

Definition 7.4.2 Loss Function. A *loss function* is a real-valued function of two variables, $L(\theta, a)$, where $\theta \in \Omega$ and a is a real number. The interpretation is that the statistician loses $L(\theta, a)$ if the parameter equals θ and the estimate equals a.

As before, let $\xi(\theta)$ denote the prior p.d.f. of θ on the set Ω, and consider a problem in which the statistician must estimate the value of θ without being able to observe the values in a random sample. If the statistician chooses a particular estimate a, then her expected loss will be

$$E[L(\theta, a)] = \int_{\Omega} L(\theta, a)\xi(\theta)\, d\theta. \tag{7.4.1}$$

We shall assume that the statistician wishes to choose an estimate a for which the expected loss in Eq. (7.4.1) is a minimum.

Definition of a Bayes Estimator

Suppose now that the statistician can observe the value $\boldsymbol{x}$ of the random vector $\boldsymbol{X}$ before estimating θ, and let $\xi(\theta|\boldsymbol{x})$ denote the posterior p.d.f. of θ on Ω. (The case of a discrete parameter can be handled in similar fashion.) For each estimate a that the statistician might use, her expected loss in this case will be

$$E[L(\theta, a)|\boldsymbol{x}] = \int_{\Omega} L(\theta, a)\xi(\theta|\boldsymbol{x})\, d\theta. \tag{7.4.2}$$

Hence, the statistician should now choose an estimate a for which the expectation in Eq. (7.4.2) is a minimum.

For each possible value $\boldsymbol{x}$ of the random vector $\boldsymbol{X}$, let $\delta^*(\boldsymbol{x})$ denote a value of the estimate a for which the expected loss in Eq. (7.4.2) is a minimum. Then the function $\delta^*(\boldsymbol{X})$ for which the values are specified in this way will be an estimator of θ.

Definition 7.4.3 Bayes Estimator/Estimate. Let $L(\theta, a)$ be a loss function. For each possible value $\boldsymbol{x}$ of $\boldsymbol{X}$, let $\delta^*(\boldsymbol{x})$ be a value of a such that $E[L(\theta, a)|\boldsymbol{x}]$ is minimized. Then δ^* is called a *Bayes estimator* of θ. Once $\boldsymbol{X} = \boldsymbol{x}$ is observed, $\delta^*(\boldsymbol{x})$ is called a *Bayes estimate* of θ.

Another way to describe a Bayes estimator δ^* is to note that, for each possible value $\boldsymbol{x}$ of $\boldsymbol{X}$, the value $\delta^*(\boldsymbol{x})$ is chosen so that

$$E[L(\theta, \delta^*(\boldsymbol{x}))|\boldsymbol{x}] = \min_{\text{All } a} E[L(\theta, a)|\boldsymbol{x}]. \tag{7.4.3}$$

In summary, we have considered an estimation problem in which a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ is to be taken from a distribution involving a parameter θ that has an unknown value in some specified set Ω. For every given loss function $L(\theta, a)$ and every prior p.d.f. $\xi(\theta)$, the Bayes estimator of θ is the estimator $\delta^*(\boldsymbol{X})$ for which Eq. (7.4.3) is satisfied for every possible value $\boldsymbol{x}$ of $\boldsymbol{X}$. It should be emphasized that the form of the Bayes estimator will depend on both the loss function that is used

in the problem and the prior distribution that is assigned to θ. In the problems described in this text, Bayes estimators will exist. However, there are more complicated situations in which no function δ^* satisfies (7.4.3).

Different Loss Functions

By far, the most commonly used loss function in estimation problems is the squared error loss function.

Definition 7.4.4 Squared Error Loss Function. The loss function

$$L(\theta, a) = (\theta - a)^2 \tag{7.4.4}$$

is called *squared error loss*.

When the squared error loss function is used, the Bayes estimate $\delta^*(\boldsymbol{x})$ for each observed value of $\boldsymbol{x}$ will be the value of a for which the expectation $E[(\theta - a)^2|\boldsymbol{x}]$ is a minimum. Theorem 4.7.3 states that, when the expectation of $(\theta - a)^2$ is calculated with respect to the posterior distribution of θ, this expectation will be a minimum when a is chosen to be equal to the mean $E(\theta|\boldsymbol{x})$ of the posterior distribution, if that posterior mean is finite. If the posterior mean of θ is not finite, then the expected loss is infinite for every possible estimate a. Hence, we have the following corollary to Theorem 4.7.3.

Corollary 7.4.1 Let θ be a real-valued parameter. Suppose that the squared error loss function (7.4.4) is used and that the posterior mean of θ, $E(\theta|\boldsymbol{X})$, is finite. Then, a Bayes estimator of θ is $\delta^*(\boldsymbol{X}) = E(\theta|\boldsymbol{X})$. ■

Example 7.4.3 Estimating the Parameter of a Bernoulli Distribution. Let the random sample $X_1, \dots, X_n$ be taken from the Bernoulli distribution with parameter θ, which is unknown and must be estimated. Let the prior distribution of θ be the beta distribution with parameters $\alpha > 0$ and $\beta > 0$. Suppose that the squared error loss function is used, as specified by Eq. (7.4.4), for $0 < \theta < 1$ and $0 < a < 1$. We shall determine the Bayes estimator of θ.

For observed values $x_1, \dots, x_n$, let $y = \sum_{i=1}^n x_i$. Then it follows from Theorem 7.3.1 that the posterior distribution of θ will be the beta distribution with parameters $\alpha_1 = \alpha + y$ and $\beta_1 = \beta + n - y$. Since the mean of the beta distribution with parameters α_1 and β_1 is $\alpha_1/(\alpha_1 + \beta_1)$, the mean of this posterior distribution of θ will be $(\alpha + y)/(\alpha + \beta + n)$. The Bayes estimate $\delta(\boldsymbol{x})$ will be equal to this value for each observed vector $\boldsymbol{x}$. Therefore, the Bayes estimator $\delta^*(\boldsymbol{X})$ is specified as follows:

$$\delta^*(\boldsymbol{X}) = \frac{\alpha + \sum_{i=1}^n X_i}{\alpha + \beta + n}. \tag{7.4.5}$$

◄

Example 7.4.4 Estimating the Mean of a Normal Distribution. Suppose that a random sample $X_1, \dots, X_n$ is to be taken from a normal distribution for which the value of the mean θ is unknown and the value of the variance σ^2 is known. Suppose also that the prior distribution of θ is the normal distribution with mean μ_0 and variance v_0^2. Suppose, finally, that the squared error loss function is to be used, as specified in Eq. (7.4.4), for $-\infty < \theta < \infty$ and $-\infty < a < \infty$. We shall determine the Bayes estimator of θ.

It follows from Theorem 7.3.3 that for all observed values $x_1, \dots, x_n$, the posterior distribution of θ will be a normal distribution with mean μ_1 specified by

Eq. (7.3.1). Therefore, the Bayes estimator $\delta^*(\boldsymbol{X})$ is specified as follows:

$$\delta^*(\boldsymbol{X}) = \frac{\sigma^2\mu_0 + nv_0^2\overline{X}_n}{\sigma^2 + nv_0^2}. \tag{7.4.6}$$

The posterior variance of θ does not enter into this calculation. ◀

Another commonly used loss function in estimation problems is the absolute error loss function.

Definition 7.4.5 Absolute Error Loss Function. The loss function

$$L(\theta, a) = |\theta - a| \tag{7.4.7}$$

is called *absolute error loss*.

For every observed value of $\boldsymbol{x}$, the Bayes estimate $\delta^*(\boldsymbol{x})$ will now be the value of a for which the expectation $E(|\theta - a||\boldsymbol{x})$ is a minimum. It was shown in Theorem 4.5.3 that for every given probability distribution of θ, the expectation of $|\theta - a|$ will be a minimum when a is chosen to be equal to a median of the distribution of θ. Therefore, when the expectation of $|\theta - a|$ is calculated with respect to the posterior distribution of θ, this expectation will be a minimum when a is chosen to be a median of the posterior distribution of θ.

Corollary 7.4.2 When the absolute error loss function (7.4.7) is used, a Bayes estimator of a real-valued parameter is $\delta^*(\boldsymbol{X})$ equal to a median of the posterior distribution of θ.

We shall now reconsider Examples 7.4.3 and 7.4.4, but we shall use the absolute error loss function instead of the squared error loss function.

Example 7.4.5 Estimating the Parameter of a Bernoulli Distribution. Consider again the conditions of Example 7.4.3, but suppose now that the absolute error loss function is used, as specified by Eq. (7.4.7). For all observed values $x_1, \ldots, x_n$, the Bayes estimate $\delta^*(\boldsymbol{x})$ will be equal to the median of the posterior distribution of θ, which is the beta distribution with parameters $\alpha + y$ and $\beta + n - y$. There is no simple expression for this median. It must be determined by numerical approximations for each given set of observed values. Most statistical computer software can compute the median of an arbitrary beta distribution.

As a specific example, consider the situation described in Example 7.3.13 in which an improper prior was used. The posterior distribution of θ in that example was the beta distribution with parameters 22 and 18. The mean of this beta distribution is $22/40 = 0.55$. The median is 0.5508. ◀

Example 7.4.6 Estimating the Mean of a Normal Distribution. Consider again the conditions of Example 7.4.4, but suppose now that the absolute error loss function is used, as specified by Eq. (7.4.7). For all observed values $x_1, \ldots, x_n$, the Bayes estimate $\delta^*(\boldsymbol{x})$ will be equal to the median of the posterior normal distribution of θ. However, since the mean and the median of each normal distribution are equal, $\delta^*(\boldsymbol{x})$ is also equal to the mean of the posterior distribution. Therefore, the Bayes estimator with respect to the absolute error loss function is the same as the Bayes estimator with respect to the squared error loss function, and it is again given by Eq. (7.4.6). ◀

Other Loss Functions Although the squared error loss function and, to a lesser extent, the absolute error loss function are the most commonly used ones in estimation problems, neither of these loss functions may be appropriate in a particular problem. In some problems, it might be appropriate to use a loss function having the form $L(\theta, a) = |\theta - a|^k$, where k is some positive number other than 1 or 2. In other problems, the loss that results when the error $|\theta - a|$ has a given magnitude might depend on the actual value of θ. In such a problem, it might be appropriate to use a loss function having the form $L(\theta, a) = \lambda(\theta)(\theta - a)^2$ or $L(\theta, a) = \lambda(\theta)|\theta - a|$, where $\lambda(\theta)$ is a given positive function of θ. In still other problems, it might be more costly to overestimate the value of θ by a certain amount than to underestimate it by the same amount. One specific loss function that reflects this property is as follows:

$$L(\theta, a) = \begin{cases} 3(\theta - a)^2 & \text{for } \theta \le a, \\ (\theta - a)^2 & \text{for } \theta > a. \end{cases}$$

Various other types of loss functions might be relevant in specific estimation problems. However, in this book we shall give most of our attention to the squared error and absolute error loss functions.

The Bayes Estimate for Large Samples

Effect of Different Prior Distributions Suppose that the proportion θ of defective items in a large shipment is unknown and that the prior distribution of θ is the uniform distribution on the interval [0, 1]. Suppose also that the value of θ must be estimated, and that the squared error loss function is used. Suppose, finally, that in a random sample of 100 items from the shipment, exactly 10 items are found to be defective. Since the uniform distribution is the beta distribution with parameters $\alpha = 1$ and $\beta = 1$, and since $n = 100$ and $y = 10$ for the given sample, it follows from Eq. (7.4.5) that the Bayes estimate is $\delta^*(\boldsymbol{x}) = 11/102 = 0.108$.

Next, suppose that the prior p.d.f. of θ has the form $\xi(\theta) = 2(1 - \theta)$ for $0 < \theta < 1$, instead of being a uniform distribution, and that again in a random sample of 100 items, exactly 10 items are found to be defective. Since $\xi(\theta)$ is the p.d.f. of the beta distribution with parameters $\alpha = 1$ and $\beta = 2$, it follows from Eq. (7.4.5) that in this case the Bayes estimate of θ is $\delta(\boldsymbol{x}) = 11/103 = 0.107$.

The two prior distributions considered here are quite different. The mean of the uniform prior distribution is 1/2, and the mean of the other beta prior distribution is 1/3. Nevertheless, because the number of observations in the sample is so large ($n = 100$), the Bayes estimates with respect to the two different prior distributions are almost the same. Furthermore, the values of both estimates are very close to the observed proportion of defective items in the sample, which is $\overline{x}_n = 0.1$.

Example 7.4.7 Chest Measurements of Scottish Soldiers. Quetelet (1846) reported (with some errors) data on the chest measurements (in inches) of 5732 Scottish militiamen. These data appeared earlier in an 1817 medical journal and are discussed by Stigler (1986). Figure 7.6 shows a histogram of the data. Suppose that we were to model the individual chest measurements as a random sample (given θ) of normal random variables with mean θ and variance 4. The average chest measurement is $\overline{x}_n = 39.85$. If θ had the normal prior distribution with mean μ_0 and variance v_0^2, then using Eq. (7.3.1) the posterior distribution of θ would be normal with mean

$$\mu_1 = \frac{4\mu_0 + 5732 \times v_0^2 \times 39.85}{4 + 5732 \times v_0^2},$$

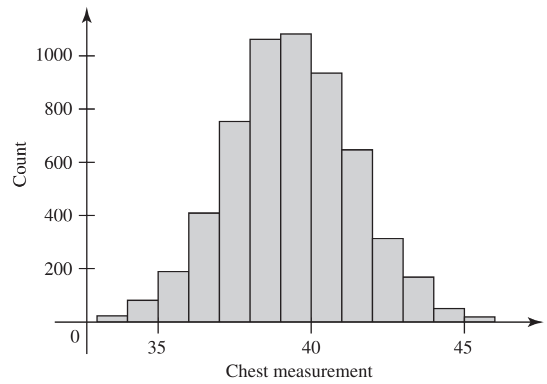


Figure 7.6 Histogram of chest measurements of Scottish militiamen in Example 7.4.7.

and variance

$$v_1^2 = \frac{4v_0^2}{4 + 5732v_0^2}.$$

The Bayes estimate will then be $\delta(\boldsymbol{x}) = \mu_1$. Notice that, unless μ_0 is incredibly large or v_0^2 is very small, we will have μ_1 nearly equal to 39.85 and v_1^2 nearly equal to $4/5732$. Indeed, if the prior p.d.f. of θ is any continuous function that is positive around $\theta = 39.85$ and is not extremely large when θ is far from 39.85, then the posterior p.d.f. of θ will very nearly be the normal p.d.f. with mean 39.85 and variance $4/5732$. The mean and median of the posterior distribution are nearly $\overline{x}_n$ regardless of the prior distribution. ◀

Consistency of the Bayes Estimator Let $X_1, \dots, X_n$ be a random sample (given θ) from the Bernoulli distribution with parameter θ. Suppose that we use a conjugate prior for θ. Since θ is the mean of the distribution from which the sample is being taken, it follows from the law of large numbers discussed in Sec. 6.2 that $\overline{X}_n$ converges in probability to θ as $n \to \infty$. Since the difference between the Bayes estimator $\delta^*(\boldsymbol{X})$ and $\overline{X}_n$ converges in probability to 0 as $n \to \infty$, it can also be concluded that $\delta^*(\boldsymbol{X})$ converges in probability to the unknown value of θ as $n \to \infty$.

Definition 7.4.6 Consistent Estimator. A sequence of estimators that converges in probability to the unknown value of the parameter being estimated, as $n \to \infty$, is called a *consistent sequence of estimators*.

Thus, we have shown that the Bayes estimators $\delta^*(\boldsymbol{X})$ form a consistent sequence of estimators in the problem considered here. The practical interpretation of this result is as follows: When large numbers of observations are taken, there is high probability that the Bayes estimator will be very close to the unknown value of θ.

The results that have just been presented for estimating the parameter of a Bernoulli distribution are also true for other estimation problems. Under fairly general conditions and for a wide class of loss functions, the Bayes estimators of some parameters θ will form a consistent sequence of estimators as the sample size $n \to \infty$. In particular, for random samples from any one of the various families of distributions discussed in Sec. 7.3, if a conjugate prior distribution is assigned to the parameters and the squared error loss function is used, the Bayes estimators will form a consistent sequence of estimators.

For example, consider again the conditions of Example 7.4.4. In that example, a random sample is taken from a normal distribution for which the value of the mean

θ is unknown, and the Bayes estimator $\delta^*(\boldsymbol{X})$ is specified by Eq. (7.4.6). By the law of large numbers, $\overline{X}_n$ will converge to the unknown value of the mean θ as $n \to \infty$. It can now be seen from Eq. (7.4.6) that $\delta^*(\boldsymbol{X})$ will also converge to θ as $n \to \infty$. Thus, the Bayes estimators again form a consistent sequence of estimators. Other examples are given in Exercises 7 and 11 at the end of this section.

More General Parameters and Estimators

So far in this section, we have considered only real-valued parameters and estimators of those parameters. There are two very common generalizations of this situation that are easy to handle with the same techniques described above. The first generalization is to multidimensional parameters such as the two-dimensional parameter of a normal distribution with unknown mean and variance. The second generalization is to functions of the parameter rather than the parameter itself. For example, if θ is the failure rate in Example 7.1.1, we might be interested in estimating $1/\theta$, the mean time to failure. As another example, if our data arise from a normal distribution with unknown mean and variance, we might wish to estimate the mean only rather than the entire parameter.

The necessary changes to Definition 7.4.1 in order to handle both of the generalizations just mentioned are given in Definition 7.4.7.

Definition 7.4.7 Estimator/Estimate. Let $X_1, \ldots, X_n$ be observable data whose joint distribution is indexed by a parameter θ taking values in a subset Ω of k-dimensional space. Let h be a function from Ω into d-dimensional space. Define $\psi = h(\theta)$. An *estimator* of ψ is a function $\delta(X_1, \ldots, X_n)$ that takes values in d-dimensional space. If $X_1 = x_1, \ldots, X_n = x_n$ are observed, then $\delta(x_1, \ldots, x_n)$ is called the *estimate* of ψ.

When h in Definition 7.4.7 is the identity function $h(\theta) = \theta$, then $\psi = \theta$ and we are estimating the original parameter θ. When $h(\theta)$ is one coordinate of θ, then the ψ that we are estimating is just that one coordinate.

There will be a number of examples of multidimensional parameters in later sections and chapters of this book. Here is an example of estimating a function of a parameter.

Example 7.4.8 Lifetimes of Electronic Components. In Example 7.3.12, suppose that we want to estimate $\psi = 1/\theta$, the mean time to failure of the electronic components. The posterior distribution of θ is the gamma distribution with parameters 4 and 8.6. If we use the squared error loss $L(\theta, a) = (\psi - a)^2$, Theorem 4.7.3 says that the Bayes estimate is the mean of the posterior distribution of ψ. That is,

$$\begin{aligned}
\delta^*(\boldsymbol{x}) = E(\psi|\boldsymbol{x}) &= E\left(\frac{1}{\theta}\,\middle|\,\boldsymbol{x}\right) \\
&= \int_0^\infty \frac{1}{\theta}\xi(\theta|\boldsymbol{x})d\theta \\
&= \int_0^\infty \frac{1}{\theta}\frac{8.6^4}{6}\theta^3 e^{-8.6\theta}d\theta \\
&= \frac{8.6^4}{6}\int_0^\infty \theta^2 e^{-8.6\theta}d\theta \\
&= \frac{8.6^4}{6}\frac{2}{8.6^3} = 2.867,
\end{aligned}$$

where the final equality follows from Theorem 5.7.3. The mean of $1/\theta$ is slightly higher than $1/E(\theta|\boldsymbol{x}) = 8.6/4 = 2.15$. ◀

Note: Loss Functions and Utility. In Sec. 4.8, we introduced the concept of utility to measure the values to a decision maker of various random outcomes. The concept of loss function is closely related to that of utility. In a sense, a loss function is like the negative of a utility. Indeed, Example 4.8.8 shows how to convert absolute error loss into a utility. In that example, Y plays the role of the parameter and $d(W)$ plays the role of the estimator. In a similar manner, one can convert other loss functions into utilities. Hence, it is not surprising that the goal of maximizing expected utility in Sec. 4.8 has been replaced by the goal of minimizing expected loss in the present section.

Limitations of Bayes Estimators

The theory of Bayes estimators, as described in this section, provides a satisfactory and coherent theory for the estimation of parameters. Indeed, according to statisticians who adhere to the Bayesian philosophy, it provides the only coherent theory of estimation that can possibly be developed. Nevertheless, there are certain limitations to the applicability of this theory in practical statistical problems. To apply the theory, it is necessary to specify a particular loss function, such as the squared error or absolute error function, and also a prior distribution for the parameter. Meaningful specifications may exist, in principle, but it may be very difficult and time-consuming to determine them. In some problems, the statistician must determine the specifications that would be appropriate for clients or employers who are unavailable or otherwise unable to communicate their preferences and knowledge. In other problems, it may be necessary for an estimate to be made jointly by members of a group or committee, and it may be difficult for the members of the group to reach agreement about an appropriate loss function and prior distribution.

Another possible difficulty is that in a particular problem the parameter θ may actually be a vector of real-valued parameters for which all the values are unknown. The theory of Bayes estimation, which has been developed in the preceding sections, can easily be generalized to include the estimation of a vector parameter θ. However, to apply this theory in such a problem it is necessary to specify a multivariate prior distribution for the vector θ and also to specify a loss function $L(\theta, \boldsymbol{a})$ that is a function of the vector θ and the vector $\boldsymbol{a}$, which will be used to estimate θ. Even though the statistician may be interested in estimating only one or two components of the vector θ in a given problem, he must still assign a multivariate prior distribution to the entire vector θ. In many important statistical problems, some of which will be discussed later in this book, θ may have a large number of components. In such a problem, it is especially difficult to specify a meaningful prior distribution on the multidimensional parameter space Ω.

It should be emphasized that there is no simple way to resolve these difficulties. Other methods of estimation that are not based on prior distributions and loss functions typically have practical limitations, also. These other methods also typically have serious defects in their theoretical structure as well.

Summary

An estimator of a parameter θ is a function δ of the data $\boldsymbol{X}$. If $\boldsymbol{X} = \boldsymbol{x}$ is observed, the value $\delta(\boldsymbol{x})$ is called our estimate, the observed value of the estimator $\delta(\boldsymbol{X})$. A loss

function $L(\theta, a)$ is designed to measure how costly it is to use the value a to estimate θ. A Bayes estimator $\delta^*(\boldsymbol{X})$ is chosen so that $a = \delta^*(\boldsymbol{x})$ provides the minimum value of the posterior mean of $L(\theta, a)$. That is,

$$E[L(\theta, \delta^*(\boldsymbol{x}))|\boldsymbol{x}] = \min_a E[L(\theta, a)|\boldsymbol{x}].$$

If the loss is squared error, $L(\theta, a) = (\theta - a)^2$, then $\delta^*(\boldsymbol{x})$ is the posterior mean of θ, $E(\theta|\boldsymbol{x})$. If the loss is absolute error, $L(\theta, a) = |\theta - a|$, then $\delta^*(\boldsymbol{x})$ is a median of the posterior distribution of θ. For other loss functions, locating the minimum might have to be done numerically.

Exercises

1. In a clinical trial, let the probability of successful outcome θ have a prior distribution that is the uniform distribution on the interval [0, 1], which is also the beta distribution with parameters 1 and 1. Suppose that the first patient has a successful outcome. Find the Bayes estimates of θ that would be obtained for both the squared error and absolute error loss functions.

2. Suppose that the proportion θ of defective items in a large shipment is unknown, and the prior distribution of θ is the beta distribution for which the parameters are $\alpha = 5$ and $\beta = 10$. Suppose also that 20 items are selected at random from the shipment, and that exactly one of these items is found to be defective. If the squared error loss function is used, what is the Bayes estimate of θ?

3. Consider again the conditions of Exercise 2. Suppose that the prior distribution of θ is as given in Exercise 2, and suppose again that 20 items are selected at random from the shipment.

- **a.** For what number of defective items in the sample will the mean squared error of the Bayes estimate be a maximum?
- **b.** For what number will the mean squared error of the Bayes estimate be a minimum?

4. Suppose that a random sample of size n is taken from the Bernoulli distribution with parameter θ, which is unknown, and that the prior distribution of θ is a beta distribution for which the mean is μ_0. Show that the mean of the posterior distribution of θ will be a weighted average having the form $\gamma_n \overline{X}_n + (1 - \gamma_n)\mu_0$, and show that $\gamma_n \to 1$ as $n \to \infty$.

5. Suppose that the number of defects in a 1200-foot roll of magnetic recording tape has a Poisson distribution for which the value of the mean θ is unknown, and the prior distribution of θ is the gamma distribution with parameters $\alpha = 3$ and $\beta = 1$. When five rolls of this tape are selected at random and inspected, the numbers of defects found on the rolls are 2, 2, 6, 0, and 3. If the squared error loss function is used, what is the Bayes estimate of θ? (See Exercise 5 of Sec. 7.3.)

6. Suppose that a random sample of size n is taken from a Poisson distribution for which the value of the mean θ is unknown, and the prior distribution of θ is a gamma distribution for which the mean is μ_0. Show that the mean of the posterior distribution of θ will be a weighted average having the form $\gamma_n \overline{X}_n + (1 - \gamma_n)\mu_0$, and show that $\gamma_n \to 1$ as $n \to \infty$.

7. Consider again the conditions of Exercise 6, and suppose that the value of θ must be estimated by using the squared error loss function. Show that the Bayes estimators, for $n = 1, 2, \ldots$, form a consistent sequence of estimators of θ.

8. Suppose that the heights of the individuals in a certain population have a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2 inches. Suppose also that the prior distribution of θ is a normal distribution for which the mean is 68 inches and the standard deviation is 1 inch. Suppose finally that 10 people are selected at random from the population, and their average height is found to be 69.5 inches.

- **a.** If the squared error loss function is used, what is the Bayes estimate of θ?
- **b.** If the absolute error loss function is used, what is the Bayes estimate of θ? (See Exercise 7 of Sec. 7.3).

9. Suppose that a random sample is to be taken from a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2, the prior distribution of θ is a normal distribution for which the standard deviation is 1, and the value of θ must be estimated by using the squared error loss function. What is the smallest random sample that must be taken in order for the mean squared error of the Bayes estimator of θ to be 0.01 or less? (See Exercise 10 of Sec. 7.3.)

10. Suppose that the time in minutes required to serve a customer at a certain facility has an exponential distribution for which the value of the parameter θ is unknown,

the prior distribution of θ is a gamma distribution for which the mean is 0.2 and the standard deviation is 1, and the average time required to serve a random sample of 20 customers is observed to be 3.8 minutes. If the squared error loss function is used, what is the Bayes estimate of θ? (See Exercise 12 of Sec. 7.3.)

11. Suppose that a random sample of size n is taken from an exponential distribution for which the value of the parameter θ is unknown, the prior distribution of θ is a specified gamma distribution, and the value of θ must be estimated by using the squared error loss function. Show that the Bayes estimators, for $n = 1, 2, \ldots$, form a consistent sequence of estimators of θ.

12. Let θ denote the proportion of registered voters in a large city who are in favor of a certain proposition. Suppose that the value of θ is unknown, and two statisticians A and B assign to θ the following different prior p.d.f.'s $\xi_A(\theta)$ and $\xi_B(\theta)$, respectively:

$$\xi_A(\theta) = 2\theta \quad \text{for } 0 < \theta < 1,$$

$$\xi_B(\theta) = 4\theta^3 \quad \text{for } 0 < \theta < 1.$$

In a random sample of 1000 registered voters from the city, it is found that 710 are in favor of the proposition.

a. Find the posterior distribution that each statistician assigns to θ.

b. Find the Bayes estimate for each statistician based on the squared error loss function.

c. Show that after the opinions of the 1000 registered voters in the random sample had been obtained, the Bayes estimates for the two statisticians could not possibly differ by more than 0.002, regardless of the number in the sample who were in favor of the proposition.

13. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where the value of the parameter θ is unknown. Suppose also that the prior distribution of θ is the Pareto distribution with parameters x_0 and α ($x_0 > 0$ and $\alpha > 0$), as defined in Exercise 16 of Sec. 5.7. If the value of θ is to be estimated by using the squared error loss function, what is the Bayes estimator of θ? (See Exercise 18 of Sec. 7.3.)

14. Suppose that $X_1, \ldots, X_n$ form a random sample from an exponential distribution for which the value of the parameter θ is unknown ($\theta > 0$). Let $\xi(\theta)$ denote the prior p.d.f. of θ, and let $\hat{\theta}$ denote the Bayes estimator of θ with respect to the prior p.d.f. $\xi(\theta)$ when the squared error loss function is used. Let $\psi = \theta^2$, and suppose that instead of estimating θ, it is desired to estimate the value of ψ subject to the following squared error loss function:

$$L(\psi, a) = (\psi - a)^2 \quad \text{for } \psi > 0 \text{ and } a > 0.$$

Let $\hat{\psi}$ denote the Bayes estimator of ψ. Explain why $\hat{\psi} > \hat{\theta}^2$. *Hint:* Look at Exercise 4 in Sec. 4.4.

15. Let $c > 0$ and consider the loss function

$$L(\theta, a) = \begin{cases} c|\theta - a| & \text{if } \theta < a, \\ |\theta - a| & \text{if } \theta \geq a. \end{cases}$$

Assume that θ has a continuous distribution. Prove that a Bayes estimator of θ will be any $1/(1 + c)$ quantile of the posterior distribution of θ. *Hint:* The proof is a lot like the proof of Theorem 4.5.3. The result holds even if θ does not have a continuous distribution, but the proof is more cumbersome.

7.5 Maximum Likelihood Estimators

Maximum likelihood estimation is a method for choosing estimators of parameters that avoids using prior distributions and loss functions. It chooses as the estimate of θ the value of θ that provides the largest value of the likelihood function.

Introduction

Example 7.5.1 Lifetimes of Electronic Components. Suppose that we observe the data in Example 7.3.11 consisting of the lifetimes of three electronic components. Is there a method for estimating the failure rate θ without first constructing a prior distribution and a loss function? ◀

In this section, we shall develop a relatively simple method of constructing an estimator without having to specify a loss function and a prior distribution. It is called the method of *maximum likelihood*, and it was introduced by R. A. Fisher in 1912. Maximum likelihood estimation can be applied in most problems, it has a strong

intuitive appeal, and it will often yield a reasonable estimator of θ. Furthermore, if the sample is large, the method will typically yield an excellent estimator of θ. For these reasons, the method of maximum likelihood is probably the most widely used method of estimation in statistics.

Note: Terminology. Because maximum likelihood estimation, as well as many other procedures to be introduced later in the text, do not involve the specification of a prior distribution of the parameter, some different terminology is often used in describing the statistical models to which these procedures are applied. Rather than saying that $X_1, \ldots, X_n$ are i.i.d. with p.f. or p.d.f. $f(x|\theta)$ conditional on θ, we might say that $X_1, \ldots, X_n$ form a random sample from a distribution with p.f. or p.d.f. $f(x|\theta)$ where θ is unknown. More specifically, in Example 7.5.1, we could say that the lifetimes form a random sample from the exponential distribution with unknown parameter θ.

Definition of a Maximum Likelihood Estimator

Let the random variables $X_1, \ldots, X_n$ form a random sample from a discrete distribution or a continuous distribution for which the p.f. or the p.d.f. is $f(x|\theta)$, where the parameter θ belongs to some parameter space Ω. Here, θ can be either a real-valued parameter or a vector. For every observed vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ in the sample, the value of the joint p.f. or joint p.d.f. will, as usual, be denoted by $f_n(\boldsymbol{x}|\theta)$. Because of its importance in this section, we repeat Definition 7.2.3.

Definition 7.5.1 Likelihood Function. When the joint p.d.f. or the joint p.f. $f_n(\boldsymbol{x}|\theta)$ of the observations in a random sample is regarded as a function of θ for given values of $x_1, \ldots, x_n$, it is called the *likelihood function*.

Consider first, the case in which the observed vector $\boldsymbol{x}$ came from a discrete distribution. If an estimate of θ must be selected, we would certainly not consider any value of $\theta \in \Omega$ for which it would be impossible to obtain the vector $\boldsymbol{x}$ that was actually observed. Furthermore, suppose that the probability $f_n(\boldsymbol{x}|\theta)$ of obtaining the actual observed vector $\boldsymbol{x}$ is very high when θ has a particular value, say, $\theta = \theta_0$, and is very small for every other value of $\theta \in \Omega$. Then we would naturally estimate the value of θ to be θ_0 (unless we had strong prior information that outweighed the evidence in the sample and pointed toward some other value). When the sample comes from a continuous distribution, it would again be natural to try to find a value of θ for which the probability density $f_n(\boldsymbol{x}|\theta)$ is large and to use this value as an estimate of θ. For each possible observed vector $\boldsymbol{x}$, we are led by this reasoning to consider a value of θ for which the likelihood function $f_n(\boldsymbol{x}|\theta)$ is a maximum and to use this value as an estimate of θ. This concept is formalized in the following definition.

Definition 7.5.2 Maximum Likelihood Estimator/Estimate. For each possible observed vector $\boldsymbol{x}$, let $\delta(\boldsymbol{x}) \in \Omega$ denote a value of $\theta \in \Omega$ for which the likelihood function $f_n(\boldsymbol{x}|\theta)$ is a maximum, and let $\hat{\theta} = \delta(\boldsymbol{X})$ be the estimator of θ defined in this way. The estimator $\hat{\theta}$ is called a *maximum likelihood estimator* of θ. After $\boldsymbol{X} = \boldsymbol{x}$ is observed, the value $\delta(\boldsymbol{x})$ is called a *maximum likelihood estimate* of θ.

The expressions *maximum likelihood estimator* and *maximum likelihood estimate* are abbreviated M.L.E. One must rely on context to determine whether the abbreviation refers to an estimator or to an estimate. Note that the M.L.E. is required to be an element of the parameter space Ω, unlike general estimators/estimates for which no such requirement exists.

Examples of Maximum Likelihood Estimators

Example 7.5.2 Lifetimes of Electronic Components. In Example 7.3.11, the observed data are $X_1 = 3$, $X_2 = 1.5$, and $X_3 = 2.1$. The random variables had been modeled as a random sample of size 3 from the exponential distribution with parameter θ. The likelihood function is, for $\theta > 0$,

$$f_3(\boldsymbol{x}|\theta) = \theta^3 \exp(-6.6\theta),$$

where $\boldsymbol{x} = (2, 1.5, 2.1)$. The value of θ that maximizes the likelihood function $f_3(\boldsymbol{x}|\theta)$ will be the same as the value of θ that maximizes $\log f_3(\boldsymbol{x}|\theta)$, since log is an increasing function. Therefore, it will be convenient to determine the M.L.E. by finding the value of θ that maximizes

$$L(\theta) = \log f_3(\boldsymbol{x}|\theta) = 3\log(\theta) - 6.6\theta.$$

Taking the derivative $dL(\theta)/d\theta$, setting the derivative to 0, and solving for θ yields $\theta = 3/6.6 = 0.455$. The second derivative is negative at this value of θ, so it provides a maximum. The maximum likelihood estimate is then 0.455. ◀

It should be noted that in some problems, for certain observed vectors $\boldsymbol{x}$, the maximum value of $f_n(\boldsymbol{x}|\theta)$ may not actually be attained for any point $\theta \in \Omega$. In such a case, an M.L.E. of θ does not exist. For certain other observed vectors $\boldsymbol{x}$, the maximum value of $f_n(\boldsymbol{x}|\theta)$ may actually be attained at more than one point in the space Ω. In such a case, the M.L.E. is not uniquely defined, and any one of these points can be chosen as the value of the estimator $\hat{\theta}$. In many practical problems, however, the M.L.E. exists and is uniquely defined.

We shall now illustrate the method of maximum likelihood and these various possibilities by considering several examples. In each example, we shall attempt to determine an M.L.E.

Example 7.5.3 Test for a Disease. Suppose that you are walking down the street and notice that the Department of Public Health is giving a free medical test for a certain disease. The test is 90 percent reliable in the following sense: If a person has the disease, there is a probability of 0.9 that the test will give a positive response; whereas, if a person does not have the disease, there is a probability of only 0.1 that the test will give a positive response. This same test was considered in Example 2.3.1. We shall let X stand for the result of the test, where $X = 1$ means that the test is positive and $X = 0$ means that the test is negative. Let the parameter space be $\Omega = \{0.1, 0.9\}$, where $\theta = 0.1$ means that the person tested does not have the disease, and $\theta = 0.9$ means that the person has the disease. This parameter space was chosen so that, given θ, X has the Bernoulli distribution with parameter θ. The likelihood function is

$$f(x|\theta) = \theta^x(1-\theta)^{1-x}.$$

If $x = 0$ is observed, then

$$f(0|\theta) = \begin{cases} 0.9 & \text{if } \theta = 0.1, \\ 0.1 & \text{if } \theta = 0.9. \end{cases}$$

Clearly, $\theta = 0.1$ maximizes the likelihood when $x = 0$ is observed. If $x = 1$ is observed, then

$$f(1|\theta) = \begin{cases} 0.1 & \text{if } \theta = 0.1, \\ 0.9 & \text{if } \theta = 0.9. \end{cases}$$

Clearly, $\theta = 0.9$ maximizes the likelihood when $x = 1$ is observed. Hence, we have that the M.L.E. is

$$\hat{\theta} = \begin{cases} 0.1 & \text{if } X = 0, \\ 0.9 & \text{if } X = 1. \end{cases}$$ ◄

Example 7.5.4

Sampling from a Bernoulli Distribution. Suppose that the random variables $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with parameter θ, which is unknown ($0 \le \theta \le 1$). For all observed values $x_1, \ldots, x_n$, where each x_i is either 0 or 1, the likelihood function is

$$f_n(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}. \tag{7.5.1}$$

Instead of maximizing the likelihood function $f_n(\boldsymbol{x}|\theta)$ directly, it is again easier to maximize $\log f_n(\boldsymbol{x}|\theta)$:

$$\begin{aligned} L(\theta) = \log f_n(\boldsymbol{x}|\theta) &= \sum_{i=1}^{n}[x_i \log\theta + (1-x_i)\log(1-\theta)] \\ &= \left(\sum_{i=1}^{n} x_i\right)\log\theta + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-\theta). \end{aligned}$$

Now calculate the derivative $dL(\theta)/d\theta$, set this derivative equal to 0, and solve the resulting equation for θ. If $\sum_{i=1}^{n} x_i \notin \{0, n\}$, we find that the derivative is 0 at $\theta = \overline{x}_n$, and it can be verified (for example, by examining the second derivative) that this value does indeed maximize $L(\theta)$ and the likelihood function defined by Eq. (7.5.1). If $\sum_{i=1}^{n} x_i = 0$, then $L(\theta)$ is a decreasing function of θ for all θ, and hence L achieves its maximum at $\theta = 0$. Similarly, if $\sum_{i=1}^{n} x_i = n$, L is an increasing function, and it achieves its maximum at $\theta = 1$. In these last two cases, note that the maximum of the likelihood occurs at $\theta = \overline{x}_n$. It follows, therefore, that the M.L.E. of θ is $\hat{\theta} = \overline{X}_n$. ◄

It follows from Example 7.5.4 that if $X_1, \ldots, X_n$ are regarded as n Bernoulli trials and if the parameter space is $\Omega = [0, 1]$, then the M.L.E. of the unknown probability of success on any given trial is simply the proportion of successes observed in the n trials. In Example 7.5.3, we have $n = 1$ Bernoulli trial, but the parameter space is $\Omega = \{0.1, 0.9\}$ rather than $[0, 1]$, and the M.L.E. differs from the proportion of successes.

Example 7.5.5

Sampling from a Normal Distribution with Unknown Mean. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which the mean μ is unknown and the variance σ^2 is known. For all observed values $x_1, \ldots, x_n$, the likelihood function $f_n(\boldsymbol{x}|\mu)$ will be

$$f_n(\boldsymbol{x}|\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]. \tag{7.5.2}$$

It can be seen from Eq. (7.5.2) that $f_n(\boldsymbol{x}|\mu)$ will be maximized by the value of μ that minimizes

$$Q(\mu) = \sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2.$$

We see that Q is a quadratic in μ with positive coefficient on μ^2. It follows that Q will be minimized where its derivative is 0. If we now calculate the derivative $dQ(\mu)/d\mu$, set this derivative equal to 0, and solve the resulting equation for μ, we find that $\mu = \overline{x}_n$. It follows, therefore, that the M.L.E. of μ is $\hat{\mu} = \overline{X}_n$. ◀

It can be seen in Example 7.5.5 that the estimator $\hat{\mu}$ is not affected by the value of the variance σ^2, which we assumed was known. The M.L.E. of the unknown mean μ is simply the sample mean $\overline{X}_n$, regardless of the value of σ^2. We shall see this again in the next example, in which both μ and σ^2 must be estimated.

Example 7.5.6 **Sampling from a Normal Distribution with Unknown Mean and Variance.** Suppose again that $X_1, \dots, X_n$ form a random sample from a normal distribution, but suppose now that both the mean μ and the variance σ^2 are unknown. The parameter is then $\theta = (\mu, \sigma^2)$. For all observed values $x_1, \dots, x_n$, the likelihood function $f_n(\boldsymbol{x}|\mu, \sigma^2)$ will again be given by the right side of Eq. (7.5.2). This function must now be maximized over all possible values of μ and σ^2, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. Instead of maximizing the likelihood function $f_n(\boldsymbol{x}|\mu, \sigma^2)$ directly, it is again easier to maximize $\log f_n(\boldsymbol{x}|\mu, \sigma^2)$. We have

$$\begin{aligned} L(\theta) &= \log f_n(\boldsymbol{x}|\mu, \sigma^2) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \end{aligned} \tag{7.5.3}$$

We shall find the value of $\theta = (\mu, \sigma^2)$ for which $L(\theta)$ is maximum in three stages. First, for each fixed σ^2, we shall find the value $\hat{\mu}(\sigma^2)$ that maximizes the right side of (7.5.3). Second, we shall find the value $\widehat{\sigma^2}$ of σ^2 that maximizes $L(\theta')$ when $\theta' = (\hat{\mu}(\sigma^2), \sigma^2)$. Finally, the M.L.E. of θ will be the random vector whose observed value is $(\hat{\mu}(\widehat{\sigma^2}), \widehat{\sigma^2})$. The first stage has already been solved in Example 7.5.5. There, we obtained $\hat{\mu}(\sigma^2) = \overline{x}_n$. For the second stage, we set $\theta' = (\overline{x}_n, \sigma^2)$ and maximize

$$L(\theta') = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2. \tag{7.5.4}$$

This can be maximized by setting its derivative with respect to σ^2 equal to 0 and solving for σ^2. The derivative is

$$\frac{d}{d\sigma^2}L(\theta') = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

Setting this to 0 yields

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2. \tag{7.5.5}$$

The second derivative of (7.5.4) is negative at the value of σ^2 in (7.5.5), so we have found the maximum. Therefore, the M.L.E. of $\theta = (\mu, \sigma^2)$ is

$$\hat{\theta} = (\hat{\mu}, \widehat{\sigma^2}) = \left(\overline{X}_n, \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right). \tag{7.5.6}$$

Notice that the first coordinate of the M.L.E. in Eq. (7.5.6) is called the sample mean of the data. Likewise, we call the second coordinate of this M.L.E. the *sample variance*. It is not difficult to see that the observed value of the sample variance is

the variance of a distribution that assigns probability $1/n$ to each of the n observed values $x_1, \dots, x_n$ in the sample. (See Exercise 1.) ◀

Example 7.5.7 Sampling from a Uniform Distribution. Suppose that $X_1, \dots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where the value of the parameter θ is unknown ($\theta > 0$). The p.d.f. $f(x|\theta)$ of each observation has the following form:

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{7.5.7}$$

Therefore, the joint p.d.f. $f_n(\boldsymbol{x}|\theta)$ of $X_1, \dots, X_n$ has the form

$$f_n(\boldsymbol{x}|\theta) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 \le x_i \le \theta \ (i = 1, \dots, n), \\ 0 & \text{otherwise.} \end{cases} \tag{7.5.8}$$

It can be seen from Eq. (7.5.8) that the M.L.E. of θ must be a value of θ for which $\theta \ge x_i$ for $i = 1, \dots, n$ and that maximizes $1/\theta^n$ among all such values. Since $1/\theta^n$ is a decreasing function of θ, the estimate will be the smallest value of θ such that $\theta \ge x_i$ for $i = 1, \dots, n$. Since this value is $\theta = \max\{x_1, \dots, x_n\}$, the M.L.E. of θ is $\hat{\theta} = \max\{X_1, \dots, X_n\}$. ◀

Limitations of Maximum Likelihood Estimation

Despite its intuitive appeal, the method of maximum likelihood is not necessarily appropriate in all problems. For instance, in Example 7.5.7, the M.L.E. $\hat{\theta}$ does not seem to be a suitable estimator of θ. Since $\max\{X_1, \dots, X_n\} < \theta$ with probability 1, it follows that $\hat{\theta}$ surely underestimates the value of θ. Indeed, if any prior distribution is assigned to θ, then the Bayes estimator of θ will surely be greater than $\hat{\theta}$. The actual amount by which the Bayes estimator exceeds $\hat{\theta}$ will, of course, depend on the particular prior distribution that is used and on the observed values of $X_1, \dots, X_n$. Example 7.5.7 also raises another difficulty with maximum likelihood, as we illustrate in Example 7.5.8.

Example 7.5.8 Nonexistence of an M.L.E. Suppose again that $X_1, \dots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$. However, suppose now that instead of writing the p.d.f. $f(x|\theta)$ of the uniform distribution in the form given in Eq. (7.5.7), we write it in the following form:

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 < x < \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{7.5.9}$$

The only difference between Eq. (7.5.7) and Eq. (7.5.9) is that the value of the p.d.f. at each of the two endpoints 0 and θ has been changed by replacing the weak inequalities in Eq. (7.5.7) with strict inequalities in Eq. (7.5.9). Therefore, either equation could be used as the p.d.f. of the uniform distribution. However, if Eq. (7.5.9) is used as the p.d.f, then an M.L.E. of θ will be a value of θ for which $\theta > x_i$ for $i = 1, \dots, n$ and which maximizes $1/\theta^n$ among all such values. It should be noted that the possible values of θ no longer include the value $\theta = \max\{x_1, \dots, x_n\}$, because θ must be *strictly* greater than each observed value x_i ($i = 1, \dots, n$). Because θ can be chosen arbitrarily close to the value $\max\{x_1, \dots, x_n\}$ but cannot be chosen equal to this value, it follows that the M.L.E. of θ does not exist. ◀

In all of our previous discussions about p.d.f.'s, we emphasized the fact that it is irrelevant whether the p.d.f. of the uniform distribution is chosen to be equal to $1/\theta$

over the open interval $0 < x < \theta$ or over the closed interval $0 \le x \le \theta$. Now, however, we see that the existence of an M.L.E. depends on this irrelevant and unimportant choice. This difficulty is easily avoided in Example 7.5.8 by using the p.d.f. given by Eq. (7.5.7) rather than that given by Eq. (7.5.9). In many other problems as well, a difficulty of this type can be avoided simply by choosing one particular appropriate version of the p.d.f. to represent the given distribution. However, as we shall see in Example 7.5.10, the difficulty cannot always be avoided.

Example 7.5.9

Non-uniqueness of an M.L.E. Suppose that $X_1, \dots, X_n$ form a random sample from the uniform distribution on the interval $[\theta, \theta + 1]$, where the value of the parameter θ is unknown ($-\infty < \theta < \infty$). In this example, the joint p.d.f. $f_n(\boldsymbol{x}|\theta)$ has the form

$$f_n(\boldsymbol{x}|\theta) = \begin{cases} 1 & \text{for } \theta \le x_i \le \theta + 1, (i = 1, \dots, n), \\ 0 & \text{otherwise.} \end{cases} \tag{7.5.10}$$

The condition that $\theta \le x_i$ for $i = 1, \dots, n$ is equivalent to the condition that $\theta \le \min\{x_1, \dots, x_n\}$. Similarly, the condition that $x_i \le \theta + 1$ for $i = 1, \dots, n$ is equivalent to the condition that $\theta \ge \max\{x_1, \dots, x_n\} - 1$. Therefore, instead of writing $f_n(\boldsymbol{x}|\theta)$ in the form given in Eq. (7.5.10), we can use the following form:

$$f_n(\boldsymbol{x}|\theta) = \begin{cases} 1 & \text{for } \max\{x_1, \dots, x_n\} - 1 \le \theta \le \min\{x_1, \dots, x_n\}, \\ 0 & \text{otherwise.} \end{cases} \tag{7.5.11}$$

Thus, it is possible to select as an M.L.E. any value of θ in the interval

$$\max\{x_1, \dots, x_n\} - 1 \le \theta \le \min\{x_1, \dots, x_n\}. \tag{7.5.12}$$

In this example, the M.L.E. is not uniquely specified. In fact, the method of maximum likelihood provides very little help in choosing an estimate of θ. The likelihood of every value of θ outside the interval (7.5.12) is actually 0. Therefore, no value θ outside this interval would ever be estimated, and all values inside the interval are M.L.E.'s. ◀

Example 7.5.10

Sampling from a Mixture of Two Distributions. Consider a random variable X that can come with equal probability either from the normal distribution with mean 0 and variance 1 or from another normal distribution with mean μ and variance σ^2, where both μ and σ^2 are unknown. Under these conditions, the p.d.f. $f(x|\mu, \sigma^2)$ of X will be the average of the p.d.f.'s of the two different normal distributions. Thus,

$$f(x|\mu, \sigma^2) = \frac{1}{2}\left\{ \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{x^2}{2}\right) + \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \right\}. \tag{7.5.13}$$

Suppose now that $X_1, \dots, X_n$ form a random sample from the distribution for which the p.d.f. is given by Eq. (7.5.13). As usual, the likelihood function $f_n(\boldsymbol{x}|\mu, \sigma^2)$ has the form

$$f_n(\boldsymbol{x}|\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i|\mu, \sigma^2). \tag{7.5.14}$$

To find the M.L.E. of $\theta = (\mu, \sigma^2)$, we must find values of μ and σ^2 for which $f_n(\boldsymbol{x}|\mu, \sigma^2)$ is maximized.

Let x_k denote any one of the observed values $x_1, \dots, x_n$. If we let $\mu = x_k$ and let $\sigma^2 \to 0$, then the factor $f(x_k|\mu, \sigma^2)$ on the right side of Eq. (7.5.14) will grow large without bound, while each factor $f(x_i|\mu, \sigma^2)$ for $x_i \ne x_k$ will approach the value

$$\frac{1}{2(2\pi)^{1/2}} \exp\left(-\frac{x_i^2}{2}\right).$$

Hence, when $\mu = x_k$ and $\sigma^2 \to 0$, we find that $f_n(\boldsymbol{x}|\mu, \sigma^2) \to \infty$.

The value 0 is not a permissible estimate of σ^2, because we know in advance that $\sigma^2 > 0$. Since the likelihood function can be made arbitrarily large by choosing $\mu = x_k$ and choosing σ^2 arbitrarily close to 0, it follows that the M.L.E. does not exist.

If we try to correct this difficulty by allowing the value 0 to be a permissible estimate of σ^2, then we find that there are n different M.L.E.'s of μ and σ^2; namely,

$$\hat{\theta}_k = (\hat{\mu}, \widehat{\sigma^2}) = (X_k, 0) \text{ for } k = 1, \ldots, n.$$

None of these estimators seems appropriate. Consider again the description, given at the beginning of this example, of the two normal distributions from which each observation might come. Suppose, for example, that $n = 1000$, and we use the estimator $\hat{\theta}_3 = (X_3, 0)$. Then, we would be estimating the value of the unknown variance to be 0; also, in effect, we would be behaving as if exactly one of the X_i's (namely, X_3) comes from the given unknown normal distribution, whereas all the other 999 observation values come from the normal distribution with mean 0 and variance 1. In fact, however, since each observation was equally likely to come from either of the two distributions, it is much more probable that hundreds of observations, rather than just one, come from the unknown normal distribution. In this example, the method of maximum likelihood is obviously unsatisfactory. A Bayesian solution to this problem is outlined in Exercise 10 in Sec. 12.5. ◀

Finally, we shall mention one point concerning the interpretation of the M.L.E. The M.L.E. is the value of θ that maximizes the conditional p.f. or p.d.f. of the data $\boldsymbol{X}$ given θ. Therefore, the maximum likelihood estimate is the value of θ that assigned the highest probability to seeing the observed data. It is not necessarily the value of the parameter that appears to be most likely given the data. To say how likely are different values of the parameter, one would need a probability distribution for the parameter. Of course, the posterior distribution of the parameter (Sec. 7.2) would serve this purpose, but no posterior distribution is involved in the calculation of the M.L.E. Hence, it is not legitimate to interpret the M.L.E. as the most likely value of the parameter after having seen the data.

For example, consider a situation covered by Example 7.5.4. Suppose that we are going to flip a coin a few times, and we are concerned with whether or not it has a slight bias toward heads or toward tails. Let $X_i = 1$ if the ith flip is heads and $X_i = 0$ if not. If we obtain four heads and one tail in the first five flips, the observed value of the M.L.E. will be 0.8. But it would be difficult to imagine a situation in which we would feel that the most likely value of θ, the probability of heads, is as large as 0.8 based on just five tosses of what appeared a priori to be a typical coin. Treating the M.L.E. as if it were the most likely value of the parameter is very much the same as ignoring the prior information about the rare disease in the medical test of Examples 2.3.1 and 2.3.3. If the test is positive in these examples, we found (in Example 7.5.3) that the M.L.E. takes the value $\hat{\theta} = 0.9$, which corresponds to having the disease. However, if the prior probability that you have the disease is as small as in Example 2.3.1, the posterior probability that you have the disease ($\theta = 0.9$) is still small even after the positive test result. The test is not accurate enough to completely overcome the prior information. So too with our coin tossing; five tosses are not enough information to overcome prior beliefs about the coin being typical. Only when the data contain much more information than is available a priori would

it be approximately correct to think of the M.L.E. as the value that we believe the parameter is most likely to be near. This could happen either when the M.L.E. is based on a lot of data or when there is very little prior information.

Summary

The maximum likelihood estimate of a parameter θ is that value of θ that provides the largest value of the likelihood function $f_n(\boldsymbol{x}|\theta)$ for fixed data $\boldsymbol{x}$. If $\delta(\boldsymbol{x})$ denotes the maximum likelihood estimate, then $\hat{\theta} = \delta(\boldsymbol{X})$ is the maximum likelihood estimator (M.L.E.). We have computed the M.L.E. when the data comprise a random sample from a Bernoulli distribution, a normal distribution with known variance, a normal distribution with both parameters unknown, or the uniform distribution on the interval $[0, \theta]$ or on the interval $[\theta, \theta + 1]$.

Exercises

1. Let $x_1, \dots, x_n$ be distinct numbers. Let Y be a discrete random variable with the following p.f.:

$$f(y) = \begin{cases} \frac{1}{n} & \text{if } y \in \{x_1, \dots, x_n\}, \\ 0 & \text{otherwise.} \end{cases}$$

Prove that $\text{Var}(Y)$ is given by Eq. (7.5.5).

2. It is not known what proportion p of the purchases of a certain brand of breakfast cereal are made by women and what proportion are made by men. In a random sample of 70 purchases of this cereal, it was found that 58 were made by women and 12 were made by men. Find the M.L.E. of p.

3. Consider again the conditions in Exercise 2, but suppose also that it is known that $\frac{1}{2} \le p \le \frac{2}{3}$. If the observations in the random sample of 70 purchases are as given in Exercise 2, what is the M.L.E. of p?

4. Suppose that $X_1, \dots, X_n$ form a random sample from the Bernoulli distribution with parameter θ, which is unknown, but it is known that θ lies in the open interval $0 < \theta < 1$. Show that the M.L.E. of θ does not exist if every observed value is 0 or if every observed value is 1.

5. Suppose that $X_1, \dots, X_n$ form a random sample from a Poisson distribution for which the mean θ is unknown, $(\theta > 0)$.

- **a.** Determine the M.L.E. of θ, assuming that at least one of the observed values is different from 0.
- **b.** Show that the M.L.E. of θ does not exist if every observed value is 0.

6. Suppose that $X_1, \dots, X_n$ form a random sample from a normal distribution for which the mean μ is known, but the variance σ^2 is unknown. Find the M.L.E. of σ^2.

7. Suppose that $X_1, \dots, X_n$ form a random sample from an exponential distribution for which the value of the parameter β is unknown $(\beta > 0)$. Find the M.L.E. of β.

8. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution for which the p.d.f. $f(x|\theta)$ is as follows:

$$f(x|\theta) = \begin{cases} e^{\theta - x} & \text{for } x > \theta, \\ 0 & \text{for } x \le \theta. \end{cases}$$

Also, suppose that the value of θ is unknown $(-\infty < \theta < \infty)$.

- **a.** Show that the M.L.E. of θ does not exist.
- **b.** Determine another version of the p.d.f. of this same distribution for which the M.L.E. of θ will exist, and find this estimator.

9. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution for which the p.d.f. $f(x|\theta)$ is as follows:

$$f(x|\theta) = \begin{cases} \theta x^{\theta - 1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Also, suppose that the value of θ is unknown $(\theta > 0)$. Find the M.L.E. of θ.

10. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution for which the p.d.f. $f(x|\theta)$ is as follows:

$$f(x|\theta) = \frac{1}{2} e^{-|x-\theta|} \quad \text{for } -\infty < x < \infty.$$

Also, suppose that the value of θ is unknown $(-\infty < \theta < \infty)$. Find the M.L.E. of θ. *Hint:* Compare this to the problem of minimizing M.A.E as in Theorem 4.5.3.

11. Suppose that $X_1, \dots, X_n$ form a random sample from the uniform distribution on the interval $[\theta_1, \theta_2]$, where both θ_1 and θ_2 are unknown ($-\infty < \theta_1 < \theta_2 < \infty$). Find the M.L.E.'s of θ_1 and θ_2.

12. Suppose that a certain large population contains k different types of individuals ($k \ge 2$), and let θ_i denote the proportion of individuals of type i, for $i = 1, \dots, k$. Here, $0 \le \theta_i \le 1$ and $\theta_1 + \cdots + \theta_k = 1$. Suppose also that in a random sample of n individuals from this population, exactly n_i individuals are of type i, where $n_1 + \cdots + n_k = n$. Find the M.L.E.'s of $\theta_1, \dots, \theta_k$.

13. Suppose that the two-dimensional vectors (X_1, Y_1), $(X_2, Y_2), \dots, (X_n, Y_n)$ form a random sample from a bivariate normal distribution for which the means of X and Y are unknown but the variances of X and Y and the correlation between X and Y are known. Find the M.L.E.'s of the means.

7.6 Properties of Maximum Likelihood Estimators

In this section, we explore several properties of M.L.E.'s, including:

- *The relationship between the M.L.E. of a parameter and the M.L.E. of a function of that parameter*
- *The need for computational algorithms*
- *The behavior of the M.L.E. as the sample size increases*
- *The lack of dependence of the M.L.E. on the sampling plan*

We also introduce a popular alternative method of estimation (method of moments) that sometimes agrees with maximum likelihood, but can sometimes be computationally simpler.

Invariance

Example 7.6.1 Lifetimes of Electronic Components. In Example 7.1.1, the parameter θ was interpreted as the failure rate of electronic components. In Example 7.4.8, we found a Bayes estimate of $\psi = 1/\theta$, the average lifetime. Is there a corresponding method for computing the M.L.E. of ψ? ◄

Suppose that $X_1, \dots, X_n$ form a random sample from a distribution for which either the p.f. or the p.d.f. is $f(x|\theta)$, where the value of the parameter θ is unknown. The parameter may be one-dimensional or a vector of parameters. Let $\hat{\theta}$ denote the M.L.E. of θ. Thus, for all observed values $x_1, \dots, x_n$, the likelihood function $f_n(\boldsymbol{x}|\theta)$ is maximized when $\theta = \hat{\theta}$.

Suppose now that we change the parameter in the distribution as follows: Instead of expressing the p.f. or the p.d.f. $f(x|\theta)$ in terms of the parameter θ, we shall express it in terms of a new parameter $\psi = g(\theta)$, where g is a one-to-one function of θ. Is there a relationship between the M.L.E. of θ and the M.L.E. of ψ?

Theorem 7.6.1 Invariance Property of M.L.E.'s. If $\hat{\theta}$ is the maximum likelihood estimator of θ and if g is a one-to-one function, then $g(\hat{\theta})$ is the maximum likelihood estimator of $g(\theta)$.

Proof The new parameter space is Γ, the image of Ω under the function g. We shall let $\theta = h(\psi)$ denote the inverse function. Then, expressed in terms of the new parameter ψ, the p.f. or p.d.f. of each observed value will be $f[x|h(\psi)]$, and the likelihood function will be $f_n[\boldsymbol{x}|h(\psi)]$.

The M.L.E. $\hat{\psi}$ of ψ will be equal to the value of ψ for which $f_n[\boldsymbol{x}|h(\psi)]$ is maximized. Since $f_n(\boldsymbol{x}|\theta)$ is maximized when $\theta = \hat{\theta}$, it follows that $f_n[\boldsymbol{x}|h(\psi)]$ is

maximized when $h(\psi) = \hat{\theta}$. Hence, the M.L.E. $\hat{\psi}$ must satisfy the relation $h(\hat{\psi}) = \hat{\theta}$ or, equivalently, $\hat{\psi} = g(\hat{\theta})$. ■

Example 7.6.2 Lifetimes of Electronic Components. According to Theorem 7.6.1, the M.L.E. of ψ is one over the M.L.E. of θ. In Example 7.5.2, we computed the observed value of $\hat{\theta} = 0.455$. The observed value of $\hat{\psi}$ would then be $1/0.455 = 2.2$. This is a bit smaller than the Bayes estimate using squared error loss of 2.867 found in Example 7.4.8. ◀

The invariance property can be extended to functions that are not one-to-one. For example, suppose that we wish to estimate the mean μ of a normal distribution when both the mean and the variance are unknown. Then μ is not a one-to-one function of the parameter $\theta = (\mu, \sigma^2)$. In this case, the function we wish to estimate is $g(\theta) = \mu$. There is a way to define the M.L.E. of a function of θ that is not necessarily one-to-one. One popular way is the following.

Definition 7.6.1 M.L.E. of a Function. Let $g(\theta)$ be an arbitrary function of the parameter, and let G be the image of Ω under the function g. For each $t \in G$, define $G_t = \{\theta : g(\theta) = t\}$ and define

$$L^*(t) = \max_{\theta \in G_t} \log f_n(\boldsymbol{x}|\theta).$$

Finally, define the M.L.E. of $g(\theta)$ to be $\hat{t}$ where

$$L^*(\hat{t}) = \max_{t \in G} L^*(t). \tag{7.6.1}$$

The following result shows how to find the M.L.E. of $g(\theta)$ based on Definition 7.6.1.

Theorem 7.6.2 Let $\hat{\theta}$ be an M.L.E. of θ, and let $g(\theta)$ be a function of θ. Then an M.L.E. of $g(\theta)$ is $g(\hat{\theta})$.

Proof We shall prove that $\hat{t} = g(\hat{\theta})$ satisfies (7.6.1). Since $L^*(t)$ is the maximum of $\log f_n(\boldsymbol{x}|\theta)$ over θ in a subset of Ω, and since $\log f_n(\boldsymbol{x}|\hat{\theta})$ is the maximum over all θ, we know that $L^*(t) \le \log f_n(\boldsymbol{x}|\hat{\theta})$ for all $t \in G$. Let $\hat{t} = g(\hat{\theta})$. We are done if we can show that $L^*(\hat{t}) = \log f_n(\boldsymbol{x}|\hat{\theta})$. Note that $\hat{\theta} \in G_{\hat{t}}$. Since $\hat{\theta}$ maximizes $f_n(\boldsymbol{x}|\theta)$ over all θ, it also maximizes $f_n(\boldsymbol{x}|\theta)$ over $\theta \in G_{\hat{t}}$. Hence, $L^*(\hat{t}) = \log f_n(\boldsymbol{x}|\hat{\theta})$ and $\hat{t} = g(\hat{\theta})$ is an M.L.E. of $g(\theta)$. ■

Example 7.6.3 Estimating the Standard Deviation and the Second Moment. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which both the mean μ and the variance σ^2 are unknown. We shall determine the M.L.E. of the standard deviation σ and the M.L.E. of the second moment of the normal distribution $E(X^2)$. It was found in Example 7.5.6 that the M.L.E. of $\theta = (\mu, \sigma^2)$ is $\hat{\theta} = (\hat{\mu}, \widehat{\sigma^2})$. From the invariance property, we can conclude that the M.L.E. $\hat{\sigma}$ of the standard deviation is simply the square root of the sample variance. In symbols, $\hat{\sigma} = (\widehat{\sigma^2})^{1/2}$. Also, since $E(X^2) = \sigma^2 + \mu^2$, the M.L.E. of $E(X^2)$ will be $\widehat{\sigma^2} + \hat{\mu}^2$. ◀

Consistency

Consider an estimation problem in which a random sample is to be taken from a distribution involving a parameter θ. Suppose that for every sufficiently large sample

size n, that is, for every value of n greater than some given minimum number, there exists a unique M.L.E. of θ. Then, under certain conditions, which are typically satisfied in practical problems, the sequence of M.L.E.'s is a consistent sequence of estimators of θ. In other words, in such problems the sequence of M.L.E.'s converges in probability to the unknown value of θ as $n \to \infty$.

We have remarked in Sec. 7.4 that under certain general conditions the sequence of Bayes estimators of a parameter θ is also a consistent sequence of estimators. Therefore, for a given prior distribution and a sufficiently large sample size n, the Bayes estimator and the M.L.E. of θ will typically be very close to each other, and both will be very close to the unknown value of θ.

We shall not present any formal details of the conditions that are needed to prove this result. (Details can be found in chapter 7 of Schervish, 1995.) We shall, however, illustrate the result by considering again a random sample $X_1, \dots, X_n$ from the Bernoulli distribution with parameter θ, which is unknown ($0 \le \theta \le 1$). It was shown in Sec. 7.4 that if the given prior distribution of θ is a beta distribution, then the difference between the Bayes estimator of θ and the sample mean $\overline{X}_n$ converges to 0 as $n \to \infty$. Furthermore, it was shown in Example 7.5.4 that the M.L.E. of θ is $\overline{X}_n$. Thus, as $n \to \infty$, the difference between the Bayes estimator and the M.L.E. will converge to 0. Finally, the law of large numbers (Theorem 6.2.4) says that the sample mean $\overline{X}_n$ converges in probability to θ as $n \to \infty$. Therefore, both the sequence of Bayes estimators and the sequence of M.L.E.'s are consistent sequences.

Numerical Computation

In many problems there exists a unique M.L.E. $\hat{\theta}$ of a given parameter θ, but this M.L.E. cannot be expressed in closed form as a function of the observations in the sample. In such a problem, for a given set of observed values, it is necessary to determine the value of $\hat{\theta}$ by numerical computation. We shall illustrate this situation by two examples.

Example 7.6.4 Sampling from a Gamma Distribution. Suppose that $X_1, \dots, X_n$ form a random sample from the gamma distribution for which the p.d.f. is as follows:

$$f(x|\alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} \quad \text{for } x > 0. \tag{7.6.2}$$

Suppose also that the value of α is unknown ($\alpha > 0$) and is to be estimated.

The likelihood function is

$$f_n(\boldsymbol{x}|\alpha) = \frac{1}{\Gamma^n(\alpha)} \left(\prod_{i=1}^{n} x_i \right)^{\alpha-1} \exp\left(- \sum_{i=1}^{n} x_i \right). \tag{7.6.3}$$

The M.L.E. of α will be the value of α that satisfies the equation

$$\frac{\partial \log f_n(\boldsymbol{x}|\alpha)}{\partial \alpha} = 0. \tag{7.6.4}$$

When we apply Eq. (7.6.4) in this example, we obtain the following equation:

$$\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \frac{1}{n} \sum_{i=1}^{n} \log x_i. \tag{7.6.5}$$

Tables of the function $\Gamma'(\alpha)/\Gamma(\alpha)$, which is called the *digamma function*, are included in various published collections of mathematical tables. The digamma function is also available in several mathematical software packages. For all given values

of $x_1, \ldots, x_n$, the unique value of α that satisfies Eq. (7.6.5) must be determined either by referring to these tables or by carrying out a numerical analysis of the digamma function. This value will be the M.L.E. of α. ◂

Example 7.6.5

Sampling from a Cauchy Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from a Cauchy distribution centered at an unknown point θ ($-\infty < \theta < \infty$), for which the p.d.f. is as follows:

$$f(x|\theta) = \frac{1}{\pi\left[1 + (x - \theta)^2\right]} \quad \text{for } -\infty < x < \infty. \tag{7.6.6}$$

Suppose also that the value of θ is to be estimated.

The likelihood function is

$$f_n(\boldsymbol{x}|\theta) = \frac{1}{\pi^n \prod_{i=1}^{n}\left[1 + (x_i - \theta)^2\right]}. \tag{7.6.7}$$

Therefore, the M.L.E. of θ will be the value that minimizes

$$\prod_{i=1}^{n}\left[1 + (x_i - \theta)^2\right]. \tag{7.6.8}$$

For most values of $x_1, \ldots, x_n$, the value of θ that minimizes the expression (7.6.8) must be determined by a numerical computation. ◂

An alternative to exact solution of Eq. (7.6.4) is to start with a heuristic estimator of α and then apply Newton's method.

Definition 7.6.2

Newton's Method. Let $f(\theta)$ be a real-valued function of a real variable, and suppose that we wish to solve the equation $f(\theta) = 0$. Let θ_0 be an initial guess at the solution. *Newton's method* replaces the initial guess with the updated guess

$$\theta_1 = \theta_0 - \frac{f(\theta_0)}{f'(\theta_0)}.$$

The rationale behind Newton's method is illustrated in Fig. 7.7. The function $f(\theta)$ is the solid curve. Newton's method approximates the curve by a line tangent to the curve, that is, the dashed line passing through the point $(\theta_0, f(\theta_0))$, indicated by the circle. The approximating line crosses the horizontal axis at the revised guess θ_1. Typically, one replaces the intial guess with the revised guess and iterates Newton's method until the results stabilize.

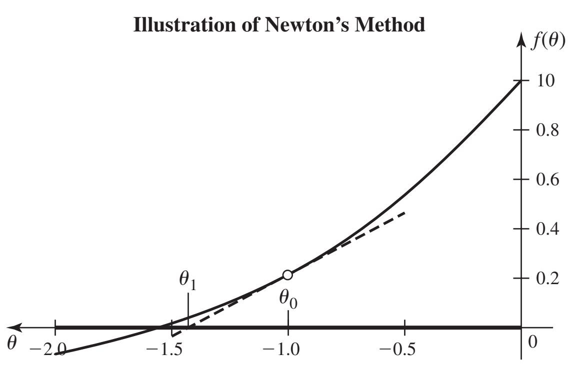


Figure 7.7 Newton's method to approximate the solution to $f(\theta) = 0$. The initial guess is θ_0, and the revised guess is θ_1.

Example 7.6.6 Sampling from a Gamma Distribution. In Example 7.6.4, suppose that we observe $n = 20$ gamma random variables $X_1, \ldots, X_{20}$ with parameters α and 1. Suppose that the observed values are such that $\frac{1}{20}\sum_{i=1}^{20}\log(x_i) = 1.220$ and $\frac{1}{20}\sum_{i=1}^{20} x_i = 3.679$. We wish to use Newton's method to approximate the M.L.E. A sensible initial guess is based on the fact that $E(X_i) = \alpha$. This suggests using $\alpha_0 = 3.679$, the sample mean. The function $f(\alpha)$ is $\psi(\alpha) - 1.220$, where ψ is the digamma function. The derivative $f'(\alpha)$ is $\psi'(\alpha)$, which is known as the trigamma function. Newton's method updates the intial guess α_0 to

$$\alpha_1 = \alpha_0 - \frac{\psi(\alpha_0) - 1.220}{\psi'(\alpha_0)} = 3.679 - \frac{1.1607 - 1.220}{0.3120} = 3.871.$$

Here, we have used statistical software that computes both the digamma and trigamma functions. After two more iterations, the approximation stabilizes at 3.876. ◀

Newton's method can fail terribly if $f'(\theta)/f(\theta)$ gets close to 0 between θ_0 and the actual solution to $f(\theta) = 0$. There is a multidimensional version of Newton's method, which we will not present here. There are also many other numerical methods for maximizing functions. Any text on numerical optimization, such as Nocedal and Wright (2006), will describe some of them.

Method of Moments

Example 7.6.7 Sampling from a Gamma Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the gamma distribution with parameters α and β. In Example 7.6.4, we explained how one could find the M.L.E. of α if β were known. The method involved the digamma function, which is unfamiliar to many people. A Bayes estimate would also be difficult to find in this example because we would have to integrate a function that includes a factor of $1/\Gamma(\alpha)^n$. Is there no other way to estimate the vector parameter θ in this example? ◀

The method of moments is an intuitive method for estimating parameters when other, more attractive, methods may be too difficult. It can also be used to obtain an initial guess for applying Newton's method.

Definition 7.6.3 Method of Moments. Assume that $X_1, \ldots, X_n$ form a random sample from a distribution that is indexed by a k-dimensional parameter θ and that has at least k finite moments. For $j = 1, \ldots, k$, let $\mu_j(\theta) = E(X_1^j|\theta)$. Suppose that the function $\mu(\theta) = (\mu_1(\theta), \ldots, \mu_k(\theta))$ is a one-to-one function of θ. Let $M(\mu_1, \ldots, \mu_k)$ denote the inverse function, that is, for all θ,

$$\theta = M(\mu_1(\theta), \ldots, \mu_k(\theta)).$$

Define the *sample moments* by $m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$ for $j = 1, \ldots, k$. The *method of moments estimator* of θ is $M(m_1, \ldots, m_j)$.

The usual way of implementing the method of moments is to set up the k equations $m_j = \mu_j(\theta)$ and then solve for θ.

Example 7.6.8 Sampling from a Gamma Distribution. In Example 7.6.4, we considered a sample of size n from the gamma distribution with parameters α and 1. The mean of each

such random variable is $\mu_1(\alpha) = \alpha$. The method of moments estimator is then $\hat{\alpha} = m_1$, the sample mean. This was the initial guess used to start Newton's method in Example 7.6.6. ◀

Example 7.6.9 Sampling from a Gamma Distribution with Both Parameters Unknown. Theorem 5.7.5 tells us that the first two moments of the gamma distribution with parameters α and β are

$$\mu_1(\theta) = \frac{\alpha}{\beta},$$
$$\mu_2(\theta) = \frac{\alpha(\alpha+1)}{\beta^2}.$$

The method of moments says to replace the right-hand sides of these equations by the sample moments and then solve for α and β. In this case, we get

$$\hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2},$$
$$\hat{\beta} = \frac{m_1}{m_2 - m_1^2}$$

as the method of moments estimators. Note that $m_2 - m_1^2$ is just the sample variance. ◀

Example 7.6.10 Sampling from a Uniform Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[\theta, \theta + 1]$, as in Example 7.5.9. In that example, we found that the M.L.E. is not unique and there is an interval of M.L.E.'s

$$\max\{x_1, \ldots, x_n\} - 1 \le \theta \le \min\{x_1, \ldots, x_n\}. \tag{7.6.9}$$

This interval contains all of the possible values of θ that are consistent with the observed data. We shall now apply the method of moments, which will produce a single estimator. The mean of each X_i is $\theta + 1/2$, so the method of moments estimator is $\overline{X}_n - 1/2$. Typically, one would expect the observed value of the method of moments estimator to be a number in the interval (7.6.9). However, that is not always the case. For example, if $n = 3$ and $X_1 = 0.2$, $X_2 = 0.99$, $X_3 = 0.01$ are observed, then (7.6.9) is the interval $[-0.01, 0.01]$, while $\overline{X}_3 = 0.4$. The method of moments estimate is then -0.1, which could not possibly be the true value of θ. ◀

There are several examples in which method of moments estimators are also M.L.E.'s. Some of these are the subjects of exercises at the end of this section.

Despite occasional problems such as Example 7.6.10, the method of moments estimators will typically be consistent in the sense of Definition 7.4.6.

Theorem 7.6.3 Suppose that $X_1, X_2, \ldots$ are i.i.d. with a distribution indexed by a k-dimensional parameter vector θ. Suppose that the first k moments of that distribution exist and are finite for all θ. Suppose also that the inverse function M in Definition 7.6.3 is continuous. Then the sequence of method of moments estimators based on $X_1, \ldots, X_n$ is a consistent sequence of estimators of θ.

Proof The law of large numbers says that the sample moments converge in probability to the moments $\mu_1(\theta), \ldots, \mu_k(\theta)$. The generalization of Theorem 6.2.5 to

functions of k variables implies that M evaluated at the sample moments (i.e., the method of moments estimator) converges in probability to θ. ■

M.L.E.'s and Bayes Estimators

Bayes estimators and M.L.E.'s depend on the data solely through the likelihood function. They use the likelihood function in different ways, but in many problems they will be very similar. When the function $f(x|\theta)$ satisfies certain smoothness conditions (as a function of θ), it can be shown that the likelihood function will tend to look more and more like a normal p.d.f. as the sample size increases. More specifically, as n increases, the likelihood function starts to look like a constant (not depending on θ, but possibly depending on the data) times

$$\exp\left[-\frac{1}{2V_n(\theta)/n}(\theta-\hat{\theta})^2\right], \tag{7.6.10}$$

where $\hat{\theta}$ is the M.L.E. and $V_n(\theta)$ is a sequence of random variables that typically converges as $n \to \infty$ to a limit that we shall call $v_\infty(\theta)$. When n is large, the function in (7.6.10) rises quickly to its peak as θ approaches $\hat{\theta}$ and then drops just as quickly as θ moves away from $\hat{\theta}$. Under these conditions, so long as the prior p.d.f. of θ is relatively flat compared to the very peaked likelihood function, the posterior p.d.f. will look a lot like the likelihood multiplied by the constant needed to turn it into a p.d.f. The posterior mean of θ will then be approximately $\hat{\theta}$. In fact, the posterior distribution of θ will be approximately the normal distribution with mean $\hat{\theta}$ and variance $V_n(\hat{\theta})/n$. In similar fashion, the distribution of the maximum likelihood estimator (given θ) will be approximately the normal distribution with mean θ and variance $v_\infty(\theta)/n$. The conditions and proofs needed to make these claims precise are beyond the scope of this text but can be found in chapter 7 of Schervish (1995).

Example 7.6.11 Sampling from an Exponential Distribution. Suppose that $X_1, X_2, \ldots$ are i.i.d. having the exponential distribution with parameter θ. Let $T_n = \sum_{i=1}^n X_i$. Then the M.L.E. of θ is $\hat{\theta}_n = n/T_n$. (This was found in Exercise 7 in Sec. 7.5.) Because $1/\hat{\theta}_n$ is an average of i.i.d. random variables with finite variance, the central limit theorem tells us that the distribution of $1/\hat{\theta}_n$ is approximately normal. The mean and variance, in this case, of that approximate normal distribution are, respectively, $1/\theta$ and $1/(\theta^2 n)$. The delta method (Theorem 6.3.2) says that $\hat{\theta}$ then has approximately the normal distribution with mean θ and variance θ^2/n. In the notation above, we have $V_n(\theta) = \theta^2$.

Next, let the prior distribution of θ be the gamma distribution with parameters α and β. Theorem 7.3.4 says that the posterior distribution of θ will be the gamma distribution with parameters $\alpha + n$ and $\beta + t_n$. We conclude by showing that this gamma distribution is approximately a normal distribution. Assume for simplicity that α is an integer. Then the posterior distribution of θ is the same as the distribution of the sum of $\alpha + n$ i.i.d. exponential random variables with parameter $\beta + t_n$. Such a sum has approximately the normal distribution with mean $(\alpha + n)/(\beta + t_n)$ and variance $(\alpha + n)/(\beta + t_n)^2$. If α and β are small, the approximate mean is then nearly $n/t_n = \hat{\theta}$, and the approximate variance is then nearly $n/t_n^2 = \hat{\theta}^2/n = V_n(\hat{\theta})/n$. ◄

Example 7.6.12 Prussian Army Deaths. In Example 7.3.14, we found the posterior distribution of θ, the mean number of deaths per year by horsekick in Prussian army units based on a sample of 280 observations. The posterior distribution was found to be the gamma distribution with parameters 196 and 280. By the same argument used in

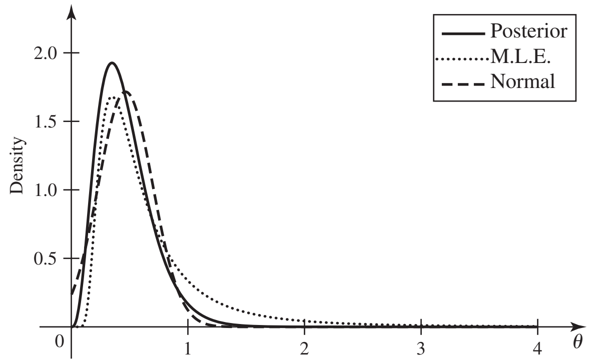


Figure 7.8 Posterior p.d.f. together with p.d.f. of M.L.E. and approximating normal p.d.f. in Example 7.6.13. For the p.d.f of the M.L.E., the value of $\theta = 3/6.6$ is used to make the p.d.f.'s as similar as possible.

Example 7.6.11, this gamma distribution is approximately the distribution of the sum of 196 i.i.d. exponential random variables with parameter 280. The distribution of this sum is approximately the normal distribution with mean 196/280 and variance $196/280^2$.

Using the same data as in Example 7.3.14, we can find the M.L.E. of θ, which is the average of the 280 observations (according to Exercise 5 in Sec. 7.5). The distribution of the average of 280 i.i.d. Poisson random variables with mean θ is approximately the normal distribution with mean θ and variance $\theta/280$ according to the central limit theorem. We then have $V_n(\theta) = \theta$ in the earlier notation. The maximum likelihood estimate with the observed data is $\hat{\theta} = 196/280$ the mean of the posterior distribution. The variance of the posterior distribution is also $V_n(\hat{\theta})/n = \hat{\theta}/280$. ◀

There are two common situations in which posterior distributions and distributions of M.L.E.'s are not such similar normal distributions as in the preceding discussion. One is when the sample size is not very large, and the other is when the likelihood function is not smooth. An example with small sample size is our electronic components example.

Example 7.6.13

Lifetimes of Electronic Components. In Example 7.3.12, we have a sample of $n = 3$ exponential random variables with parameter θ. The posterior distribution found there was the gamma distribution with parameters 4 and 8.6. The M.L.E. is $\hat{\theta} = 3/(X_1 + X_2 + X_3)$, which has the distribution of 1 over a gamma random variable with parameters 3 and 3θ. Figure 7.8 shows the posterior p.d.f. along with the p.d.f. of the M.L.E. assuming that $\theta = 3/6.6$, the observed value of the M.L.E. The two p.d.f.'s, although similar, are still different. Also, both p.d.f.'s are similar to, but still different from, the normal p.d.f. with the same mean and variance as the posterior, which also appears on the plot. ◀

An example of an unsmooth likelihood function involves the uniform distribution on the interval $[0, \theta]$.

Example 7.6.14

Sampling from a Uniform Distribution. In Example 7.5.7, we found the M.L.E. of θ based on a sample of size n from the uniform distribution on the interval $[0, \theta]$. The M.L.E. is $\hat{\theta} = \max\{X_1, \ldots, X_n\}$. We can find the exact distribution of $\hat{\theta}$ using the result in Example 3.9.6. The p.d.f. of $Y = \hat{\theta}$ is

$$g_n(y|\theta) = n[F(y|\theta)]^{n-1} f(y|\theta), \qquad (7.6.11)$$

where $f(\cdot|\theta)$ is the p.d.f. of the uniform distribution on $[0, \theta]$ and $F(\cdot|\theta)$ is the corresponding c.d.f. Substituting these well-known functions into Eq. (7.6.11) yields the p.d.f. of $Y = \hat{\theta}$:

$$g_n(y|\theta) = n\left[\frac{y}{\theta}\right]^{n-1}\frac{1}{\theta} = n\frac{y^{n-1}}{\theta^n},$$

for $0 < y < \theta$. This p.d.f. is not the least bit like a normal p.d.f. It is very asymmetric and has its maximum at the largest possible value of the M.L.E. In fact, one can compute the mean and variance of $\hat{\theta}$, respectively, as

$$E(\hat{\theta}) = \frac{n}{n+1}\theta,$$
$$Var(\hat{\theta}) = \frac{n}{(n+1)^2(n+2)}\theta^2.$$

The variance goes down like $1/n^2$ instead of like $1/n$ in the approximately normal examples we saw earlier.

If n is large, the posterior distribution of θ will have a p.d.f. that is approximately the likelihood function times the constant needed to make it into a p.d.f. The likelihood is in Eq. (7.5.8). Integrating that function over θ to obtain the needed constant leads to the following approximate posterior p.d.f. of θ:

$$\xi(\theta|\boldsymbol{x}) \approx \begin{cases} \dfrac{(n-1)\hat{\theta}^{n-1}}{\theta^n} & \text{for } \theta > \hat{\theta}, \\ 0 & \text{otherwise.} \end{cases}$$

The mean and variance of this approximate posterior distribution are, respectively, $(n-1)\hat{\theta}/(n-2)$ and $(n-1)\hat{\theta}^2/[(n-2)^2(n-3)]$. The posterior mean is still nearly equal to the M.L.E. (but a little larger), and the posterior variance decreases at a rate like $1/n^2$, as does the variance of the M.L.E. But the posterior distribution is not the least bit normal, as the p.d.f. has its maximum at the smallest possible value of θ and decreases from there. ◀

The EM Algorithm

There are a number of complicated situations in which it is difficult to compute the M.L.E. Many of these situations involve forms of missing data. The term "missing data" can refer to several different types of information. The most obvious would be observations that we had planned or hoped to observe but were not observed. For example, imagine that we planned to collect both heights and weights for a sample of athletes. For reasons that might be beyond our control, it is possible that we observed both heights and weights for most of the athletes, but only heights for one subset of atheletes and only weights for another subset. If we model the heights and weights as having a bivariate normal distribution, we might want to compute the M.L.E. of the parameters of that distribution. For a complete collection of pairs, Exercise 24 in this section gives formulas for the M.L.E. It is not difficult to see how much more complicated it would be to compute the M.L.E. in the situation described above with missing data.

The *EM algorithm* is an iterative method for approximating M.L.E.'s when missing data are making it difficult to find the M.L.E.'s in closed form. One begins (as in most iterative procedures) at stage 0 with an initial parameter vector $\theta^{(0)}$. To move from stage j to stage $j+1$, one first writes the *full-data log-likelihood*, which is what the logarithm of the likelihood function would be if we had observed the

missing data. The values of the missing data appear in the full-data log-likelihood as random variables rather than as observed values. The "E" step of the EM algorithm is the following: Compute the conditional distribution of the missing data given the observed data as if the parameter θ were equal to $\theta^{(j)}$, and then compute the conditional mean of the full-data log-likelihood treating θ as constant and the missing data as random variables. The E step gets rid of the unobserved random variables from the full-data log-likelihood and leaves θ where it was. For the "M" step, choose $\theta^{(j+1)}$ to maximize the expected value of the full-data log-likelihood that you just computed. The M step takes you to stage $j+1$. Ideally, the maximization step is no harder than it would be if the missing data had actually been observed.

Example 7.6.15

Heights and Weights. Suppose that we try to observe $n = 6$ pairs of heights and weights, but we get only three complete vectors plus one lone weight and two lone heights. We model the pairs as bivariate normal random vectors, and we want to find the M.L.E. of the parameter vector $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. (This example is for illustrative purposes. One cannot expect to get a good estimate of a five-dimensional parameter vector with only nine observed values and no prior information.) The data are in Table 7.1. The missing weights are $X_{4,2}$ and $X_{5,2}$. The missing height is $X_{6,1}$. The full-data log-likelihood is the sum of the logarithms of six expressions of the form Eq. (5.10.2) each with one of the rows of Table 7.1 substituted for the dummy variables (x_1, x_2). For example, the term corresponding to the fourth row of Table 7.1 is

$$-\log(2\pi\sigma_1\sigma_2) - \frac{1}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}\left[\left(\frac{68-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{68-\mu_1}{\sigma_1}\right)\left(\frac{X_{4,2}-\mu_2}{\sigma_2}\right) + \left(\frac{X_{4,2}-\mu_2}{\sigma_2}\right)^2\right]. \tag{7.6.12}$$

As an initial parameter vector we choose a naïve estimate computed from the observed data:

$$\theta^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, \sigma_1^{2(0)}, \sigma_2^{2(0)}, \rho^{(0)}) = (69.60, 194.75, 2.87, 14.82, 0.1764).$$

This consists of the M.L.E.'s based on the marginal distributions of the two coordinates, together with the sample correlation computed from the three complete observations.

Table 7.1 Heights and weights for Example 7.6.15. The missing values are given random variable names.

Height	Weight
72	197
70	204
73	208
68	$X_{4,2}$
65	$X_{5,2}$
$X_{6,1}$	170

The E step pretends that $\theta = \theta^{(0)}$ and computes the conditional mean of the full-data log-likelihood given the observed data. For the fourth row of Table 7.1, the conditional distribution of $X_{4,2}$ given the observed data and $\theta = \theta^{(0)}$ can be found from Theorem 5.10.4 to be the normal distribution with mean

$$194.75 + 0.1764 \times (14.82)^{1/2} \left(\frac{68 - 69.60}{2.87^{1/2}} \right) = 193.3$$

and variance $(1 - 0.1764^2)14.82^2 = 212.8$. The conditional mean of $(X_{4,2} - \mu_2)^2$ would then be $212.8 + (193.3 - \mu_2)^2$. The conditional mean of the expression in (7.6.12) would then be

$$-\log(2\pi\sigma_1\sigma_2) - \frac{1}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)} \left[\left(\frac{68 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{68 - \mu_1}{\sigma_1} \right) \left(\frac{193.3 - \mu_2}{\sigma_2} \right) + \left(\frac{193.3 - \mu_2}{\sigma_2} \right)^2 + \frac{212.8}{\sigma_2^2} \right].$$

The point to notice about this last expression is that, except for the last term $212.8/\sigma_2^2$, it is exactly the contribution to the log-likelihood that we would have obtained if $X_{4,2}$ had been observed to equal 193.3, its conditional mean. Similar calculations can be done for the other two observations with missing coordinates. Each will produce a contribution to the log-likelihood that is the conditional variance of the missing coordinate divided by its variance plus what the log-likelihood would have been if the missing value had been observed to equal its conditional mean. This makes the M step almost identical to finding the M.L.E. for a completely observed data set. The only difference from the formulas in Exercise 24 is the following: For each observation that is missing X, add the conditional variance of X given Y to $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ in both the formula for $\widehat{\sigma_1^2}$ and $\hat{\rho}$. Similarly, for each observation that is missing Y, add the conditional variance of Y given X to $\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$ in both the formula for $\widehat{\sigma_2^2}$ and $\hat{\rho}$.

We now illustrate the first iteration of the EM algorithm with the data of this example. We already have $\theta^{(0)}$, and we can compute the log-likelihood function from the observed data at $\theta^{(0)}$ as -31.359. To begin the algorithm, we have already computed the conditional mean and variance of the missing second coordinate from the fourth row of Table 7.1. The corresponding conditional means and variances for the fifth and sixth rows are 190.6 and 212.8 for the fifth row and 68.76 and 7.98 for the sixth row. For the E step, we replace the missing observations by their conditional means and add the conditional variances to the sums of squared deviations. For the M step, we insert the values just computed into the formulas of Exercise 24 as described above. The new vector is

$$\theta^{(1)} = (69.46, 193.81, 2.88, 14.83, 0.3742),$$

and the log-likelihood is -31.03. After 32 iterations, the estimate and log-likelihood stop changing. The final estmate is

$$\theta^{(32)} = (68.86, 189.71, 3.15, 15.03, 0.8965),$$

with log-likelihood -29.66. ◂

Example 7.6.16 Mixture of Normal Distributions. A very popular use of the EM algorithm is in fitting mixture distributions. Let $X_1, \ldots, X_n$ be random variables such that each one is

sampled either from the normal distribution with mean μ_1 and variance σ^2 (with probability p) or from the normal distribution with mean μ_2 and variance σ^2 (with probability $1-p$), where $\mu_1 < \mu_2$. The restriction that $\mu_1 < \mu_2$ is to make the model identifiable in the following sense. If $\mu_1 = \mu_2$ is allowed, then every value of p leads to the same joint distribution of the observable data. Also, if neither mean is constrained to be below the other, then switching the two means and changing p to $1-p$ will produce the same joint distribution for the observable data. The restriction $\mu_1 < \mu_2$ ensures that every distinct parameter vector produces a different joint distribution for the observable data.

The data in Fig. 7.4 have the typical appearance of a distribution that is a mixture of two normals with means not very far apart. Because we have assumed that the variances of the two distributions are the same, we will not have the problem that arose in Example 7.5.10.

The likelihood function from observations $X_1 = x_1, \dots, X_n = x_n$ is

$$\prod_{i=1}^{n}\left[\frac{p}{(2\pi)^{1/2}\sigma}\exp\left(\frac{-(x_i-\mu_1)^2}{2\sigma^2}\right)+\frac{1-p}{(2\pi)^{1/2}\sigma}\exp\left(\frac{-(x_i-\mu_2)^2}{2\sigma^2}\right)\right]. \quad (7.6.13)$$

The parameter vector is $\theta = (\mu_1, \mu_2, \sigma^2, p)$, and maximizing the likelihood as written is a challenge. However, we can introduce missing observations $Y_1, \dots, Y_n$ where $Y_i = 1$ if X_i was sampled from the distribution with mean μ_1 and $Y_i = 0$ if X_i was sampled from the distribution with mean μ_2. The full-data log-likelihood can be written as the sum of the logarithm of the marginal p.f. of the missing Y data plus the logarithm of the conditional p.d.f. of the observed X data given the Y data. That is,

$$\begin{aligned}\sum_{i=1}^{n} Y_i \log(p) + \left(n - \sum_{i=1}^{n} Y_i\right)\log(1-p) - \frac{n}{2}\log(2\pi\sigma^2)\\ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i(x_i-\mu_1)^2 + (1-Y_i)(x_i-\mu_2)^2\right].\end{aligned} \quad (7.6.14)$$

At stage j with estimate $\theta^{(j)}$ of θ, the E step first finds the conditional distribution of $Y_1, \dots, Y_n$ given the observed data and $\theta = \theta^{(j)}$. Since $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent pairs, we can find the conditional distribution separately for each pair. The joint distribution of (X_i, Y_i) is a mixed distribution with p.f./p.d.f.

$$f(x_i, y_i|\theta^{(j)}) = \frac{p^{y_i}(1-p)^{1-y_i}}{(2\pi)^{1/2}\sigma^{(j)}}\exp\left(-\frac{1}{\sigma^{2(j)}}\left[y_i(x_i-\mu_1^{(j)})^2 + (1-y_i)(x_i-\mu_2^{(j)})^2\right]\right).$$

The marginal p.d.f. of X_i is the ith factor in (7.6.13). It is straightforward to determine that the conditional distribution of Y_i given the observed data is the Bernoulli distribution with parameter

$$q_i^{(j)} = \frac{p^{(j)}\exp\left(-\frac{(x_i-\mu_1^{(j)})^2}{2\sigma^{2(j)}}\right)}{p^{(j)}\exp\left(-\frac{(x_i-\mu_1^{(j)})^2}{2\sigma^{2(j)}}\right) + (1-p^{(j)})\exp\left(-\frac{(x_i-\mu_2^{(j)})^2}{2\sigma^{2(j)}}\right)}. \quad (7.6.15)$$

Because the full-data log-likelihood is a linear function of the Y_i's, the E step simply replaces each Y_i in (7.6.14) by $q_i^{(j)}$. The result is

$$\sum_{i=1}^{n} q_i^{(j)} \log(p) + \left(n - \sum_{i=1}^{n} q_i^{(j)}\right) \log(1-p) - \frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[q_i^{(j)}(x_i-\mu_1)^2 + (1-q_i^{(j)})(x_i-\mu_2)^2\right]. \tag{7.6.16}$$

Maximizing (7.6.16) is straightforward. Since p appears in only the first two terms, we see that $p^{(j+1)}$ is just the average of the $q_i^{(j)}$'s. Also, $\mu_1^{(j+1)}$ is the weighted average of the X_i's with weights $q_i^{(j)}$. Similarly, $\mu_2^{(j+1)}$ is the weighted average of the X_i's with weights $1-q_i^{(j)}$. Finally,

$$\sigma^{2(j+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[q_i^{(j)}(x_i-\mu_1^{(j+1)})^2 + (1-q_i^{(j)})(x_i-\mu_2^{(j+1)})^2\right]. \tag{7.6.17}$$

We will illustrate the first E and M steps using the data in Example 7.3.10. For the initial parameter vector $\theta^{(0)}$, we will let $\mu_1^{(0)}$ be the average of the 10 lowest observations and $\mu_2^{(0)}$ be the average of the 10 highest observations. We set $p^{(0)} = 1/2$, and $\sigma^{2(0)}$ is the average of the sample variance of the 10 lowest observations and the sample variance of the 10 highest observations. This makes

$$\theta^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, \sigma^{2(0)}, p^{(0)}) = (-7.65, 7.36, 46.28, 0.5).$$

For each of the 20 observed values x_i, we compute $q_i^{(0)}$. For example, $x_{10} = -4.0$. According to (7.6.15),

$$q_{10}^{(0)} = \frac{0.5\exp\left(-\frac{(-4.0+7.65)^2}{2\times 46.28}\right)}{0.5\exp\left(-\frac{(-4.0+7.65)^2}{2\times 46.28}\right) + 0.5\exp\left(-\frac{(-4.0-7.36)^2}{2\times 46.28}\right)} = 0.7774.$$

A similar calculation for $x_8 = 9.0$ yields $q_8^{(0)} = 0.0489$. The initial log-likelihood, calculated as the logarithm of (7.6.13), is -75.98. The average of the 20 $q_i^{(0)}$ values is $p^{(1)} = 0.4402$. The weighted average of the data values using the $q_i^{(0)}$'s as weights is $\mu_1^{(1)} = -7.736$, and the weighted average using the $1-q_i^{(0)}$'s is $\mu_2^{(1)} = 6.3068$. Using (7.6.17), we get $\sigma^{2(1)} = 56.5491$. The log-likelihood rises to -75.19. After 25 iterations, the results settle on $\theta^{(25)} = (-21.9715, 2.6802, 48.6864, 0.1037)$ with a final log-likelihood of -72.84. The histogram from Fig. 7.4 is reproduced in Fig. 7.9 together with the p.d.f. of an observation from the fitted mixture distribution, namely,

$$f(x) = \frac{0.1037}{(2\pi\times 48.6864)^{1/2}}\exp\left(-\frac{(x+21.9715)^2}{2\times 48.6864}\right) + \frac{1-0.1037}{(2\pi\times 48.6864)^{1/2}}\exp\left(-\frac{(x-2.6802)^2}{2\times 48.6864}\right).$$

In addition, the fitted p.d.f. based on a single normal distribution is also shown in Fig. 7.9. The mean and variance of that single normal distribution are 0.1250 and 110.6809, respectively. ◀

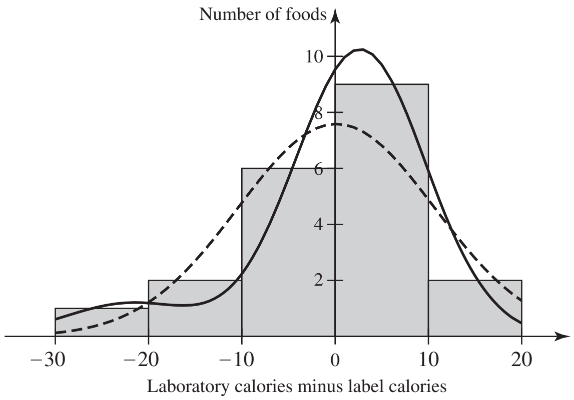


Figure 7.9 Histogram of data from Example 7.3.10 together with fitted p.d.f. from Example 7.6.16 (solid curve). The p.d.f. has been scaled up to match the fact that the histogram gives counts rather than an estimated p.d.f. Also, the dashed curve gives the estimated p.d.f. for a single normal distribution.

One can prove that the log-likelihood increases with each iteration of the EM algorithm and that the algorithm converges to a local maximum of the likelihood function. As with other numerical maximization routines, it is difficult to guarantee convergence to a global maximum.

Sampling Plans

Suppose that an experimenter wishes to take observations from a distribution for which the p.f. or the p.d.f. is $f(x|\theta)$ in order to gain information about the value of the parameter θ. The experimenter could simply take a random sample of a predetermined size from the distribution. Instead, however, he may begin by first observing a few values at random from the distribution and noting the cost and the time spent in taking these observations. He may then decide to observe a few more values at random from the distribution and to study all the values thus far obtained. At some point, the experimenter will decide to stop taking observations and will estimate the value of θ from all the observed values that have been obtained up to that point. He might decide to stop because either he feels that he has enough information to be able to make a good estimate of θ or he cannot afford to spend any more money or time on sampling.

In this experiment, the number n of observations in the sample is not fixed beforehand. It is a random variable whose value may very well depend on the magnitudes of the observations as they are obtained.

Suppose that an experimenter contemplates using a sampling plan in which, for every n, the decision of whether or not to stop sampling after n observations have been collected is a function of the n observations seen so far. Regardless of whether the experimenter chooses such a sampling plan or decides to fix the value of n before

any observations are taken, it can be shown that the likelihood function based on the observed values is proportional (as a function of θ) to

$$f(x_1|\theta) \dots f(x_n|\theta).$$

In such a situation, the M.L.E. of θ will depend only on the likelihood function and not on what type of sampling plan is used. In other words, the value of $\hat{\theta}$ depends only on the values $x_1, \dots, x_n$ that are actually observed and does not depend on the plan (if there was one) that was used by the experimenter to decide when to stop sampling.

To illustrate this property, suppose that the intervals of time, in minutes, between arrivals of successive customers at a certain service facility are i.i.d. random variables. Suppose also that each interval has the exponential distribution with parameter θ, and that a set of observed intervals $X_1, \dots, X_n$ form a random sample from this distribution. It follows from Exercise 7 of Sec. 7.5 that the M.L.E. of θ will be $\hat{\theta} = 1/\overline{X}_n$. Also, since the mean μ of the exponential distribution is $1/\theta$, it follows from the invariance property of M.L.E.'s that $\hat{\mu} = \overline{X}_n$. In other words, the M.L.E. of the mean is the average of the observations in the sample.

Consider now the following three sampling plans:

1. An experimenter decides in advance to take exactly 20 observations, and the average of these 20 observations turns out to be 6. Then the M.L.E. of μ is $\hat{\mu} = 6$.
2. An experimenter decides to take observations $X_1, X_2 \dots$ until she obtains a value greater than 10. She finds that $X_i < 10$ for $i = 1, \dots, 19$ and that $X_{20} > 10$. Hence, sampling terminates after 20 observations. If the average of these 20 observations is 6, then the M.L.E. is again $\hat{\mu} = 6$.
3. An experimenter takes observations one at a time, with no particular plan in mind, until either she is forced to stop sampling or she gets tired of sampling. She is certain that neither of these causes (being forced to stop or getting tired) depends in any way on μ. If for either reason she stops as soon as she has taken 20 observations and if the average of the 20 observations is 6, then the M.L.E. is again $\hat{\mu} = 6$.

Sometimes, an experiment of this type must be terminated during an interval when the experimenter is waiting for the next customer to arrive. If a certain amount of time has elapsed since the arrival of the last customer, this time should not be omitted from the sample data, even though the full interval to the arrival of the next customer has not been observed. Suppose, for example, that the average of the first 20 observations is 6, the experimenter waits another 15 minutes but no other customer arrives, and then she terminates the experiment. In this case, we know that the M.L.E. of μ would have to be greater than 6, since the value of the 21st observation must be greater than 15, even though its exact value is unknown. The new M.L.E. can be obtained by multiplying the likelihood function for the first 20 observations by the probability that the 21st observation is greater than 15, namely, $\exp(-15\theta)$, and finding the value of θ that maximizes this new likelihood function (see Exercise 15).

Remember that the M.L.E. is determined by the likelihood function. The only way in which the M.L.E. is allowed to depend on the sampling plan is through the likelihood function. If the decision about when to stop observing data is based solely on the observations seen so far, then this information has already been included in the likelihood function. If the decision to stop is based on something else, one needs

to evaluate the probability of that "something else" given each possible value of θ and include that probability in the likelihood.

Other properties of M.L.E.'s will be discussed later in this chapter and in Chapter 8.

Summary

The M.L.E. of a function $g(\theta)$ is $g(\hat{\theta})$, where $\hat{\theta}$ is the M.L.E. of θ. For example, if θ is the rate at which customers are served in a queue, then $1/\theta$ is the average service time. The M.L.E. of $1/\theta$ is 1 over the M.L.E. of θ. Sometimes we cannot find a closed form expression for the M.L.E. of a parameter and we must resort to numerical methods to find or approximate the M.L.E. In most problems, the sequence of M.L.E.'s, as sample size increases, converges in probability to the parameter. When data are collected in such a way that the decision to stop collecting data is based solely on the data already observed or on other considerations that are not related to the parameter, then the M.L.E. will not depend on the sampling plan. That is, if two different sampling plans lead to proportional likelihood functions, then the value of θ that maximizes one likelihood will also maximize the other.

Exercises

1. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution with the p.d.f. given in Exercise 10 of Sec. 7.5. Find the M.L.E. of $e^{-1/\theta}$.

2. Suppose that $X_1, \ldots, X_n$ form a random sample from a Poisson distribution for which the mean is unknown. Determine the M.L.E. of the standard deviation of the distribution.

3. Suppose that $X_1, \ldots, X_n$ form a random sample from an exponential distribution for which the value of the parameter β is unknown. Determine the M.L.E. of the median of the distribution.

4. Suppose that the lifetime of a certain type of lamp has an exponential distribution for which the value of the parameter β is unknown. A random sample of n lamps of this type are tested for a period of T hours and the number X of lamps that fail during this period is observed, but the times at which the failures occurred are not noted. Determine the M.L.E. of β based on the observed value of X.

5. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[a, b]$, where both endpoints a and b are unknown. Find the M.L.E. of the mean of the distribution.

6. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which both the mean and the variance are unknown. Find the M.L.E. of the 0.95 quantile of the distribution, that is, of the point θ such that $\Pr(X < \theta) = 0.95$.

7. For the conditions of Exercise 6, find the M.L.E. of $\nu = \Pr(X > 2)$.

8. Suppose that $X_1, \ldots, X_n$ form a random sample from a gamma distribution for which the p.d.f. is given by Eq. (7.6.2). Find the M.L.E. of $\Gamma'(\alpha)/\Gamma(\alpha)$.

9. Suppose that $X_1, \ldots, X_n$ form a random sample from a gamma distribution for which both parameters α and β are unknown. Find the M.L.E. of α/β.

10. Suppose that $X_1, \ldots, X_n$ form a random sample from a beta distribution for which both parameters α and β are unknown. Show that the M.L.E.'s of α and β satisfy the following equation:

$$\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} - \frac{\Gamma'(\hat{\beta})}{\Gamma(\hat{\beta})} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{X_i}{1 - X_i}.$$

11. Suppose that $X_1, \ldots, X_n$ form a random sample of size n from the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown. Show that the sequence of M.L.E.'s of θ is a consistent sequence.

12. Suppose that $X_1, \ldots, X_n$ form a random sample from an exponential distribution for which the value of the parameter β is unknown. Show that the sequence of M.L.E.'s of β is a consistent sequence.

13. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. is as specified in Exercise 9 of Section 7.5. Show that the sequence of M.L.E.'s of θ is a consistent sequence.

14. Suppose that a scientist desires to estimate the proportion p of monarch butterflies that have a special type of marking on their wings.

a. Suppose that he captures monarch butterflies one at a time until he has found five that have this special marking. If he must capture a total of 43 butterflies, what is the M.L.E. of p?

b. Suppose that at the end of a day the scientist had captured 58 monarch butterflies and had found only three with the special marking. What is the M.L.E. of p?

15. Suppose that 21 observations are taken at random from an exponential distribution for which the mean μ is unknown ($\mu > 0$), the average of 20 of these observations is 6, and although the exact value of the other observation could not be determined, it was known to be greater than 15. Determine the M.L.E. of μ.

16. Suppose that each of two statisticians A and B must estimate a certain parameter θ whose value is unknown ($\theta > 0$). Statistician A can observe the value of a random variable X, which has the gamma distribution with parameters α and β, where $\alpha = 3$ and $\beta = \theta$; statistician B can observe the value of a random variable Y, which has the Poisson distribution with mean 2θ. Suppose that the value observed by statistician A is $X = 2$ and the value observed by statistician B is $Y = 3$. Show that the likelihood functions determined by these observed values are proportional, and find the common value of the M.L.E. of θ obtained by each statistician.

17. Suppose that each of two statisticians A and B must estimate a certain parameter p whose value is unknown ($0 < p < 1$). Statistician A can observe the value of a random variable X, which has the binomial distribution with parameters $n = 10$ and p; statistician B can observe the value of a random variable Y, which has the negative binomial distribution with parameters $r = 4$ and p. Suppose that the value observed by statistician A is $X = 4$ and the value observed by statistician B is $Y = 6$. Show that the likelihood functions determined by these observed values are proportional, and find the common value of the M.L.E. of p obtained by each statistician.

18. Prove that the method of moments estimator for the parameter of a Bernoulli distribution is the M.L.E.

19. Prove that the method of moments estimator for the parameter of an exponential distribution is the M.L.E.

20. Prove that the method of moments estimator of the mean of a Poisson distribution is the M.L.E.

21. Prove that the method of moments estimators of the mean and variance of a normal distribution are also the M.L.E.'s.

22. Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution on the interval $[0, \theta]$.

a. Find the method of moments estimator of θ.

b. Show that the method of moments estimator is not the M.L.E.

23. Suppose that $X_1, \ldots, X_n$ form a random sample from the beta distribution with parameters α and β. Let $\theta = (\alpha, \beta)$ be the vector parameter.

a. Find the method of moments estimator for θ.

b. Show that the method of moments estimator is not the M.L.E.

24. Suppose that the two-dimensional vectors (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ form a random sample from a bivariate normal distribution for which the means of X and Y, the variances of X and Y, and the correlation between X and Y are unknown. Show that the M.L.E.'s of these five parameters are as follows:

$$\hat{\mu}_1 = \overline{X}_n \quad \text{and} \quad \hat{\mu}_2 = \overline{Y}_n,$$

$$\widehat{\sigma_1^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad \text{and} \quad \widehat{\sigma_2^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2,$$

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\left[\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right]^{1/2}\left[\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2\right]^{1/2}}.$$

Hint: First, rewrite the joint p.d.f. of each pair (X_i, Y_i) as the product of the marginal p.d.f. of X_i and the conditional p.d.f. of Y_i given X_i. Second, transform the parameters to μ_1, σ_1^2 and

$$\alpha = \mu_2 - \frac{\rho\sigma_2\mu_1}{\sigma_1},$$

$$\beta = \frac{\rho\sigma_2}{\sigma_1},$$

$$\sigma_{2.1}^2 = (1 - \rho^2)\sigma_2^2.$$

Third, maximize the likelihood function as a function of the new parameters. Finally, apply the invariance property of M.L.E.'s to find the M.L.E.'s of the original parameters. The above transformation greatly simplifies the maximization of the likelihood.

25. Consider again the situation described in Exercise 24. This time, suppose that, for reasons unrelated to the values of the parameters, we cannot observe the values of $Y_{n-k+1}, \ldots, Y_n$. That is, we will be able to observe all of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_{n-k}$, but not the last k Y values. Using the hint given in Exercise 24, find the M.L.E.'s of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and ρ.

★ 7.7 Sufficient Statistics

In the first six sections of this chapter, we presented some inference methods that are based on the posterior distribution of the parameter or on the likelihood function alone. There are other inference methods that are based neither on the posterior distribution nor on the likelihood function. These methods are based on the conditional distributions of various functions of the data (i.e., statistics) given the parameter. There are many statistics available in a given problem, some more useful than others. Sufficient statistics turn out to be the most useful in some sense.

Definition of a Sufficient Statistic

Example 7.7.1 Lifetimes of Electronic Components. In Examples 7.4.8 and 7.5.2, we computed estimates of the mean lifetime for electronic components based on a sample of size three from the distribution of lifetimes. The two estimates we computed were a Bayes estimate (Example 7.4.8) and an M.L.E. (Example 7.5.2). Both estimates made use of the observed data solely through the value of the statistic $X_1 + X_2 + X_3$. Is there anything special about this statistic, and if so, do such statistics exist in other problems? ◀

In many problems in which a parameter θ must be estimated, it is possible to find either an M.L.E. or a Bayes estimator that will be suitable. In some problems, however, neither of these estimators may be suitable or available. There may not be any M.L.E., or there may be more than one. Even when an M.L.E. is unique, it may not be a suitable estimator, as in Example 7.5.7, where the M.L.E. always underestimates the value of θ. Reasons why there may not be a suitable Bayes estimator were presented at the end of Sec. 7.4. In such problems, the search for a good estimator must be extended beyond the methods that have been introduced thus far. In this section, we shall define the concept of a sufficient statistic, which was introduced by R. A. Fisher in 1922, and we shall show how this concept can be used to simplify the search for a good estimator in many problems.

Suppose that in a specific estimation problem, two statisticians A and B must estimate the value of the parameter θ. Statistician A can observe the values of the observations $X_1, \ldots, X_n$ in a random sample, and statistician B cannot observe the individual values of $X_1, \ldots, X_n$ but can learn the value of a certain statistic $T = r(X_1, \ldots, X_n)$. In this case, statistician A can choose any function of the observations $X_1, \ldots, X_n$ as an estimator of θ (including a function of T). But statistician B can use only a function of T. Hence, it follows that A will generally be able to find a better estimator than will B.

In some problems, however, B will be able to do just as well as A. In such a problem, the single function $T = r(X_1, \ldots, X_n)$ will in some sense summarize all the information contained in the random sample, and knowledge of the individual values of $X_1, \ldots, X_n$ will be irrelevant in the search for a good estimator of θ. A statistic T having this property is called a *sufficient statistic*. The formal definition of a sufficient statistic is based on the following intuition. Suppose that one could learn T and were then able to simulate random variables $X'_1, \ldots, X'_n$ such that, for every θ, the joint distribution of $X'_1, \ldots, X'_n$ was exactly the same as the joint distribution of $X_1, \ldots, X_n$. Such a statistic T is sufficient in the sense that one could, if one felt the need, use $X'_1, \ldots, X'_n$ in the same way that one would have used $X_1, \ldots, X_n$. The process of simulating $X'_1, \ldots, X'_n$ is called an *auxiliary randomization*.

Definition 7.7.1 Sufficient Statistic. Let $X_1, \ldots, X_n$ be a random sample from a distribution indexed by a parameter θ. Let T be a statistic. Suppose that, for every θ and every possible value t of T, the conditional joint distribution of $X_1, \ldots, X_n$ given that $T = t$ (and θ) depends only on t but not on θ. That is, for each t, the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ and θ is the same for all θ. Then we say that T is a *sufficient statistic for the parameter* θ.

Return now to the intuition introduced right before Definition 7.7.1. When one simulates $X'_1, \ldots, X'_n$ in accordance with the conditional joint distribution of $X_1, \ldots, X_n$ given $T = t$, it follows that for each given value of $\theta \in \Omega$, the joint distribution of $T, X'_1, \ldots, X'_n$ will be the same as the joint distribution of $T, X_1, \ldots, X_n$. By integrating out (or summing out) T from the joint distribution, we see that the joint distribution of $X_1, \ldots, X_n$ is the same as the joint distribution of $X'_1, \ldots, X'_n$. Hence, if statistician B can observe the value of a sufficient statistic T, then she can generate n random variables $X'_1, \ldots, X'_n$, which have the same joint distribution as the original random sample $X_1, \ldots, X_n$. The property that distinguishes a sufficient statistic T from a statistic that is not sufficient may be described as follows: The auxiliary randomization used to generate the random variables $X'_1, \ldots, X'_n$ after the sufficient statistic T has been observed does not require any knowledge about the value of θ, since the conditional joint distribution of $X_1, \ldots, X_n$ when T is given does not depend on the value of θ. If the statistic T were not sufficient, this auxiliary randomization could not be carried out, because the conditional joint distribution of $X_1, \ldots, X_n$ for a given value of T would involve the value of θ, and this value is unknown.

If statistician B is concerned solely with the distribution of the estimator she uses, we can now see why she can estimate θ just as well as can statistician A, who observes the values of $X_1, \ldots, X_n$. Suppose that A plans to use a particular estimator $\delta(X_1, \ldots, X_n)$ to estimate θ, and B observes the value of T and generates $X'_1, \ldots, X'_n$, which have the same joint distribution as $X_1, \ldots, X_n$. If B uses the estimator $\delta(X'_1, \ldots, X'_n)$, then it follows that the probability distribution of B's estimator will be the same as the probability distribution of A's estimator. This discussion illustrates why, when searching for a good estimator, a statistician can restrict the search to estimators that are functions of a sufficient statistic T. We shall return to this point in Sec. 7.9.

On the other hand, if statistician B is interested in basing her estimator on the posterior distribution of θ, we have not yet shown why she can do just as well as statistician A. The next result (the factorization criterion) shows why even this is true. A sufficient statistic is sufficient for being able to compute the likelihood function, and hence it is sufficient for performing any inference that depends on the data only through the likelihood function. M.L.E.'s and anything based on posterior distributions depend on the data only through the likelihood function.

The Factorization Criterion

Immediately after Example 7.2.7 and Theorems 7.3.2 and 7.3.3, we pointed out that a particular statistic was used to compute the posterior distribution being discussed. These statistics all had the property that they were all that was needed from the data to be able to compute the likelihood function. This property is another way to characterize sufficient statistics. We shall now present a simple method for finding a sufficient statistic that can be applied in many problems. This method is based on the following result, which was developed with increasing generality by R. A. Fisher in 1922, J. Neyman in 1935, and P. R. Halmos and L. J. Savage in 1949.

Theorem 7.7.1 Factorization Criterion. Let $X_1, \dots, X_n$ form a random sample from either a continuous distribution or a discrete distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where the value of θ is unknown and belongs to a given parameter space Ω. A statistic $T = r(X_1, \dots, X_n)$ is a sufficient statistic for θ if and only if the joint p.d.f. or the joint p.f. $f_n(\boldsymbol{x}|\theta)$ of $X_1, \dots, X_n$ can be factored as follows for all values of $\boldsymbol{x} = (x_1, \dots, x_n) \in R^n$ and all values of $\theta \in \Omega$:

$$f_n(\boldsymbol{x}|\theta) = u(\boldsymbol{x})v[r(\boldsymbol{x}), \theta]. \tag{7.7.1}$$

Here, the functions u and v are nonnegative, the function u may depend on $\boldsymbol{x}$ but does not depend on θ, and the function v will depend on θ but depends on the observed value $\boldsymbol{x}$ only through the value of the statistic $r(\boldsymbol{x})$.

Proof We shall give the proof only when the random vector $\boldsymbol{X} = (X_1, \dots, X_n)$ has a discrete distribution, in which case

$$f_n(\boldsymbol{x}|\theta) = \Pr(\boldsymbol{X} = \boldsymbol{x}|\theta).$$

Suppose first that $f_n(\boldsymbol{x}|\theta)$ can be factored as in Eq. (7.7.1) for all values of $\boldsymbol{x} \in \boldsymbol{R}^n$ and $\theta \in \Omega$. For each possible value t of T, let $A(t)$ denote the set of all points $\boldsymbol{x} \in \boldsymbol{R}^n$ such that $r(\boldsymbol{x}) = t$. For each given value of $\theta \in \Omega$, we shall determine the conditional distribution of $\boldsymbol{X}$ given that $T = t$. For every point $\boldsymbol{x} \in A(t)$,

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) = \frac{\Pr(\boldsymbol{X} = \boldsymbol{x}|\theta)}{\Pr(T = t|\theta)} = \frac{f_n(\boldsymbol{x}|\theta)}{\sum_{\boldsymbol{y}\in A(t)} f_n(\boldsymbol{y}|\theta)}.$$

Since $r(\boldsymbol{y}) = t$ for every point $\boldsymbol{y} \in A(t)$, and since $\boldsymbol{x} \in A(t)$, it follows from Eq. (7.7.1) that

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) = \frac{u(\boldsymbol{x})}{\sum_{\boldsymbol{y}\in A(t)} u(\boldsymbol{y})}. \tag{7.7.2}$$

Finally, for every point $\boldsymbol{x}$ that does not belong to $A(t)$,

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) = 0. \tag{7.7.3}$$

It can be seen from Eqs. (7.7.2) and (7.7.3) that the conditional distribution of $\boldsymbol{X}$ does not depend on θ. Therefore, T is a sufficient statistic.

Conversely, suppose that T is a sufficient statistic. Then, for every given value t of T, every point $\boldsymbol{x} \in A(t)$, and every value of $\theta \in \Omega$, the conditional probability $\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta)$ will not depend on θ and will therefore have the form

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) = u(\boldsymbol{x}).$$

If we let $v(t, \theta) = \Pr(T = t|\theta)$, it follows that

$$\begin{aligned} f_n(\boldsymbol{x}|\theta) &= \Pr(\boldsymbol{X} = \boldsymbol{x}|\theta) = \Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) \Pr(T = t|\theta) \\ &= u(\boldsymbol{x})v(t, \theta). \end{aligned}$$

Hence, $f_n(\boldsymbol{x}|\theta)$ has been factored in the form specified in Eq. (7.7.1).

The proof for a random sample $X_1, \dots, X_n$ from a continuous distribution requires somewhat different methods and will not be given here. ■

One way to read Theorem 7.7.1 is that $T = r(\boldsymbol{X})$ is sufficient if and only if the likelihood function is proportional (as a function of θ) to a function that depends on the data only through $r(\boldsymbol{x})$. That function would be $v[r(\boldsymbol{x}), \theta]$. When using the likelihood function for finding posterior distributions, we saw that any factor not depending on θ (such as $u(\boldsymbol{x})$ in Eq. (7.7.1)) can be removed from the likelihood without affecting

the calculation of the posterior distribution. So, we have the following corollary to Theorem 7.7.1.

Corollary 7.7.1 A statistic $T = r(\boldsymbol{X})$ is sufficient if and only if, no matter what prior distribution we use, the posterior distribution of θ depends on the data only through the value of T. ■

For each value of $\boldsymbol{x}$ for which $f_n(\boldsymbol{x}|\theta) = 0$ for all values of $\theta \in \Omega$, the value of the function $u(\boldsymbol{x})$ in Eq. (7.7.1) can be chosen to be 0. Therefore, when the factorization criterion is being applied, it is sufficient to verify that a factorization of the form given in Eq. (7.7.1) is satisfied for every value of $\boldsymbol{x}$ such that $f_n(\boldsymbol{x}|\theta) > 0$ for at least one value of $\theta \in \Omega$.

We shall now illustrate the use of the factorization criterion by giving four examples.

Example 7.7.2 **Sampling from a Poisson Distribution.** Suppose that $\boldsymbol{X} = (X_1, \dots, X_n)$ form a random sample from a Poisson distribution for which the value of the mean θ is unknown ($\theta > 0$). Let $r(\boldsymbol{x}) = \sum_{i=1}^n x_i$. We shall show that $T = r(\boldsymbol{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.

For every set of nonnegative integers $x_1, \dots, x_n$, the joint p.f. $f_n(\boldsymbol{x}|\theta)$ of $X_1, \dots, X_n$ is as follows:

$$f_n(\boldsymbol{x}|\theta) = \prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \left(\prod_{i=1}^n \frac{1}{x_i!}\right) e^{-n\theta}\theta^{r(\boldsymbol{x})}.$$

Let $u(\boldsymbol{x}) = \prod_{i=1}^n (1/x_i!)$ and $v(t, \theta) = e^{-n\theta}\theta^t$. We now see that $f_n(\boldsymbol{x}|\theta)$ has been factored as in Eq. (7.7.1). It follows that $T = \sum_{i=1}^n X_i$ is a sufficient statistic for θ. ◄

Example 7.7.3 **Applying the Factorization Criterion to a Continuous Distribution.** Suppose that $\boldsymbol{X} = (X_1, \dots, X_n)$ form a random sample from a continuous distribution with the following p.d.f.:

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

It is assumed that the value of the parameter θ is unknown ($\theta > 0$). Let $r(\boldsymbol{x}) = \prod_{i=1}^n x_i$. We shall show that $T = r(\boldsymbol{X}) = \prod_{i=1}^n X_i$ is a sufficient statistic for θ.

For $0 < x_i < 1$ ($i = 1, \dots, n$), the joint p.d.f. $f_n(\boldsymbol{x}|\theta)$ of $X_1, \dots, X_n$ is as follows:

$$f(\boldsymbol{x}|\theta) = \theta^n \left(\prod_{i=1}^n x_i\right)^{\theta-1} = \theta^n \left[r(\boldsymbol{x})\right]^{\theta-1}. \tag{7.7.4}$$

Furthermore, if at least one value of x_i is outside the interval $0 < x_i < 1$, then $f_n(\boldsymbol{x}|\theta) = 0$ for every value of $\theta \in \Omega$. The right side of Eq. (7.7.4) depends on $\boldsymbol{x}$ only through the value of $r(\boldsymbol{x})$. Therefore, if we let $u(\boldsymbol{x}) = 1$ and $v(t, \theta) = \theta^n t^{\theta-1}$, then $f_n(\boldsymbol{x}|\theta)$ in Eq. (7.7.4) can be considered to be factored in the form specified in Eq. (7.7.1). It follows from the factorization criterion that the statistic $T = \prod_{i=1}^n X_i$ is a sufficient statistic for θ. ◄

Example 7.7.4 **Sampling from a Normal Distribution.** Suppose that $\boldsymbol{X} = (X_1, \dots, X_n)$ form a random sample from a normal distribution for which the mean μ is unknown and the variance σ^2 is known. Let $r(\boldsymbol{x}) = \sum_{i=1}^n x_i$. We shall show that $T = r(\boldsymbol{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for μ.

For $-\infty < x_i < \infty$ $(i = 1, \ldots, n)$, the joint p.d.f. of $\boldsymbol{X}$ is as follows:

$$f_n(\boldsymbol{x}|\mu) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]. \tag{7.7.5}$$

This equation can be rewritten in the form

$$f_n(\boldsymbol{x}|\mu) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right) \exp\left(\frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2}\right). \tag{7.7.6}$$

Let $u(\boldsymbol{x})$ be the constant factor and the first exponential factor in Eq. (7.7.6). Let $v(t, \mu) = \exp(\mu t/\sigma^2 - n\mu^2/\sigma^2)$. Then $f_n(\boldsymbol{x}|\mu)$ has now been factored as in Eq. (7.7.1). It follows from the factorization criterion that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for μ. ◀

Since $\sum_{i=1}^{n} x_i = n\overline{x}_n$, we can state equivalently that the final factor in Eq. (7.7.6) depends on $x_1, \ldots, x_n$ only through the value of $\overline{x}_n$. Therefore, in Example 7.7.4 the statistic $\overline{X}_n$ is also a sufficient statistic for μ. More generally (see Exercise 13 at the end of this section), every one-to-one function of a sufficient statistic is also a sufficient statistic.

Example 7.7.5 Sampling from a Uniform Distribution. Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where the value of the parameter θ is unknown $(\theta > 0)$. Let $r(\boldsymbol{x}) = \max\{x_1, \ldots, x_n\}$. We shall show that $T = r(\boldsymbol{X}) = \max\{X_1, \ldots, X_n\}$ is a sufficient statistic for θ.

The p.d.f. $f(x|\theta)$ of each individual observation X_i is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the joint p.d.f. $f_n(\boldsymbol{x}|\theta)$ of $X_1, \ldots, X_n$ is

$$f_n(\boldsymbol{x}|\theta) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 \le x_i \le \theta, (i = 1, \ldots, n), \\ 0 & \text{otherwise.} \end{cases}$$

It can be seen that if $x_i < 0$ for at least one value of i $(i = 1, \ldots, n)$, then $f_n(\boldsymbol{x}|\theta) = 0$ for every value of $\theta > 0$. Therefore, it is only necessary to consider the factorization of $f_n(\boldsymbol{x}|\theta)$ for values of $x_i \ge 0$ $(i = 1, \ldots, n)$.

Let $v[t, \theta]$ be defined as follows:

$$v[t, \theta] = \begin{cases} \frac{1}{\theta^n} & \text{if } t \le \theta, \\ 0 & \text{if } t > \theta. \end{cases}$$

Notice that $x_i \le \theta$ for $i = 1, \ldots, n$ if and only if $\max\{x_1, \ldots, x_n\} \le \theta$. Therefore, for $x_i \ge 0$ $(i = 1, \ldots, n)$, we can rewrite $f_n(\boldsymbol{x}|\theta)$ as follows:

$$f_n(\boldsymbol{x}|\theta) = v[r(\boldsymbol{x}), \theta]. \tag{7.7.7}$$

Letting $u(\boldsymbol{x}) = 1$, we see that the right side of Eq. (7.7.7) is in the form of Eq. (7.7.1). It follows that $T = \max\{X_1, \ldots, X_n\}$ is a sufficient statistic for θ. ◀

Summary

A statistic $T = r(\boldsymbol{X})$ is sufficient if, for each t, the conditional distribution of $\boldsymbol{X}$ given $T = t$ and θ is the same for all values of θ. So, if T is sufficient, and one observed only T instead of $\boldsymbol{X}$, one could, at least in principle, simulate random variables $\boldsymbol{X}'$ with

the same joint distribution given θ as $\boldsymbol{X}$. In this sense, T is sufficient for obtaining as much information about θ as one could get from $\boldsymbol{X}$. The factorization criterion says that $T = r(\boldsymbol{X})$ is sufficient if and only if the joint p.f. or p.d.f. can be factored as $f(\boldsymbol{x}|\theta) = u(\boldsymbol{x})v[r(\boldsymbol{x}), \theta]$ for some functions u and v. This is the most convenient way to identify whether or not a statistic is sufficient.

Exercises

Instructions for Exercises 1 to 10: In each of these exercises, assume that the random variables $X_1, \dots, X_n$ form a random sample of size n from the distribution specified in that exercise, and show that the statistic T specified in the exercise is a sufficient statistic for the parameter.

1. The Bernoulli distribution with parameter p, which is unknown $(0 < p < 1)$; $T = \sum_{i=1}^{n} X_i$.

2. The geometric distribution with parameter p, which is unknown $(0 < p < 1)$; $T = \sum_{i=1}^{n} X_i$.

3. The negative binomial distribution with parameters r and p, where r is known and p is unknown $(0 < p < 1)$; $T = \sum_{i=1}^{n} X_i$.

4. The normal distribution for which the mean μ is known and the variance $\sigma^2 > 0$ is unknown; $T = \sum_{i=1}^{n}(X_i - \mu)^2$.

5. The gamma distribution with parameters α and β, where the value of α is known and the value of β is unknown $(\beta > 0)$; $T = \overline{X}_n$.

6. The gamma distribution with parameters α and β, where the value of β is known and the value of α is unknown $(\alpha > 0)$; $T = \prod_{i=1}^{n} X_i$.

7. The beta distribution with parameters α and β, where the value of β is known and the value of α is unknown $(\alpha > 0)$; $T = \prod_{i=1}^{n} X_i$.

8. The uniform distribution on the integers $1, 2, \dots, \theta$, as defined in Sec. 3.1, where the value of θ is unknown $(\theta = 1, 2, \dots)$; $T = \max\{X_1, \dots, X_n\}$.

9. The uniform distribution on the interval $[a, b]$, where the value of a is known and the value of b is unknown $(b > a)$; $T = \max\{X_1, \dots, X_n\}$.

10. The uniform distribution on the interval $[a, b]$, where the value of b is known and the value of a is unknown $(a < b)$; $T = \min\{X_1, \dots, X_n\}$.

11. Assume that $X_1, \dots, X_n$ form a random sample from a distribution that belongs to an exponential family of distributions as defined in Exercise 23 of Sec. 7.3. Prove that $T = \sum_{i=1}^{n} d(X_i)$ is a sufficient statistic for θ.

12. Suppose that a random sample $X_1, \dots, X_n$ is drawn from the Pareto distribution with parameters x_0 and α. (See Exercise 16 in Sec. 5.7.)

- **a.** If x_0 is known and $\alpha > 0$ unknown, find a sufficient statistic.
- **b.** If α is known and x_0 unknown, find a sufficient statistic.

13. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution for which the p.d.f. is $f(x|\theta)$, where the value of the parameter θ belongs to a given parameter space Ω. Suppose that $T = r(X_1, \dots, X_n)$ and $T' = r'(X_1, \dots, X_n)$ are two statistics such that T' is a one-to-one function of T; that is, the value of T' can be determined from the value of T without knowing the values of $X_1, \dots, X_n$, and the value of T can be determined from the value of T' without knowing the values of $X_1, \dots, X_n$. Show that T' is a sufficient statistic for θ if and only if T is a sufficient statistic for θ.

14. Suppose that $X_1, \dots, X_n$ form a random sample from the gamma distribution specified in Exercise 6. Show that the statistic $T = \sum_{i=1}^{n} \log X_i$ is a sufficient statistic for the parameter α.

15. Suppose that $X_1, \dots, X_n$ form a random sample from the beta distribution with parameters α and β, where the value of α is known and the value of β is unknown $(\beta > 0)$. Show that the following statistic T is a sufficient statistic for β:

$$T = \frac{1}{n}\left(\sum_{i=1}^{n} \log \frac{1}{1 - X_i}\right)^4.$$

16. Let θ be a parameter with parameter space Ω equal to an interval of real numbers (possibly unbounded). Let $\boldsymbol{X}$ have p.d.f. or p.f. $f_n(\boldsymbol{x}|\theta)$ conditional on θ. Let $T = r(\boldsymbol{X})$ be a statistic. Assume that T is sufficient. Prove that, for every possible prior p.d.f. for θ, the posterior p.d.f. of θ given $\boldsymbol{X} = \boldsymbol{x}$ depends on $\boldsymbol{x}$ only through $r(\boldsymbol{x})$.

17. Let θ be a parameter, and let $\boldsymbol{X}$ be discrete with p.f. $f_n(\boldsymbol{x}|\theta)$ conditional on θ. Let $T = r(\boldsymbol{X})$ be a statistic. Prove that T is sufficient if and only if, for every t and every $\boldsymbol{x}$ such that $t = r(\boldsymbol{x})$, the likelihood function from observing $T = t$ is proportional to the likelihood function from observing $\boldsymbol{X} = \boldsymbol{x}$.

⋆ 7.8 Jointly Sufficient Statistics

When a parameter θ is multidimensional, sufficient statistics will typically need to be multidimensional as well. Sometimes, no one-dimensional statistic is sufficient even when θ is one-dimensional. In either case, we need to extend the concept of sufficient statistic to deal with cases in which more than one statistic is needed in order to be sufficient.

Definition of Jointly Sufficient Statistics

Example 7.8.1 Sampling from a Normal Distribution. Return to Example 7.7.4, in which $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from the normal distribution with mean μ and variance σ^2. This time, assume that both coordinates of the parameter $\theta = (\mu, \sigma^2)$ are unknown. The joint p.d.f. of $\boldsymbol{X}$ is still given by the right side of Eq. (7.7.5). But now, we would refer to the joint p.d.f. as $f_n(\boldsymbol{x}|\theta)$. With both μ and σ^2 unknown, there no longer appears to be a single statistic that is sufficient. ◄

We shall continue to suppose that the variables $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where the parameter θ must belong to some parameter space Ω. However, we shall now explicitly consider the possibility that θ may be a vector of real-valued parameters. For example, if the sample comes from a normal distribution for which both the mean μ and the variance σ^2 are unknown, then θ would be a two-dimensional vector whose components are μ and σ^2. Similarly, if the sample comes from a uniform distribution on some interval $[a, b]$ for which both endpoints a and b are unknown, then θ would be a two-dimensional vector whose components are a and b. We shall, of course, continue to include the possibility that θ is a one-dimensional parameter.

In almost every problem in which θ is a vector, as well as in some problems in which θ is one-dimensional, there does not exist a one-dimensional statistic T that is sufficient. In such a problem it is necessary to find two or more statistics $T_1, \ldots, T_k$ that together are *jointly sufficient statistics* in a sense that will now be described.

Suppose that in a given problem the statistics $T_1, \ldots, T_k$ are defined by k different functions of the vector of observations $\boldsymbol{X} = (X_1, \ldots, X_n)$. Specifically, let $T_i = r_i(\boldsymbol{X})$ for $i = 1, \ldots, k$. Loosely speaking, the statistics $T_1, \ldots, T_k$ are jointly sufficient statistics for θ if a statistician who learns only the values of the k functions $r_1(\boldsymbol{X}), \ldots, r_k(\boldsymbol{X})$ can estimate every component of θ and every function of the components of θ, as well as one who observes the n individual values of $X_1, \ldots, X_n$. More formally, we have the following definition.

Definition 7.8.1 Jointly Sufficient Statistics. Suppose that for each θ and each possible value $(t_1, \ldots, t_k)$ of $(T_1, \ldots, T_k)$, the conditional joint distribution of $(X_1, \ldots, X_n)$ given $(T_1, \ldots, T_k) = (t_1, \ldots, t_k)$ does not depend on θ. Then $T_1, \ldots, T_k$ are called *jointly sufficient statistics for θ.*

A version of the factorization criterion exists for jointly sufficient statistics. The proof will not be given, but it is similar to the proof of Theorem 7.7.1.

Theorem 7.8.1 Factorization Criterion for Jointly Sufficient Statistics. Let $r_1, \ldots, r_k$ be functions of n real variables. The statistics $T_i = r_i(\boldsymbol{X})$, $i = 1, \ldots, k$, are jointly sufficient statistics for θ if and only if the joint p.d.f. or the joint p.f. $f_n(\boldsymbol{x}|\theta)$ can be factored as follows for

all values of $\boldsymbol{x} \in \boldsymbol{R}^n$ and all values of $\theta \in \Omega$:

$$f_n(\boldsymbol{x}|\theta) = u(\boldsymbol{x})v[r_1(\boldsymbol{x}), \ldots, r_k(\boldsymbol{x}), \theta]. \tag{7.8.1}$$

Here the functions u and v are nonnegative, the function u may depend on $\boldsymbol{x}$ but does not depend on θ, and the function v will depend on θ but depends on $\boldsymbol{x}$ only through the k functions $r_1(\boldsymbol{x}), \ldots, r_k(\boldsymbol{x})$. ∎

Example 7.8.2 Jointly Sufficient Statistics for the Parameters of a Normal Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which both the mean μ and the variance σ^2 are unknown. The joint p.d.f. of $X_1, \ldots, X_n$ is given by Eq. (7.7.6), and it can be seen that this joint p.d.f. depends on $\boldsymbol{x}$ only through the values of $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i^2$. Therefore, by the factorization criterion, the statistics $T_1 = \sum_{i=1}^n X_i$ and $T_2 = \sum_{i=1}^n X_i^2$ are jointly sufficient statistics for μ and σ^2. ◀

Suppose now that in a given problem the statistics $T_1, \ldots, T_k$ are jointly sufficient statistics for some parameter vector θ. If k other statistics $T_1', \ldots, T_k'$ are obtained from $T_1, \ldots, T_k$ by a one-to-one transformation, then it can be shown that $T_1', \ldots, T_k'$ will also be jointly sufficient statistics for θ.

Example 7.8.3 Another Pair of Jointly Sufficient Statistics for the Parameters of a Normal Distribution. Suppose again that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which both the mean μ and the variance σ^2 are unknown. Let $T_1' = \hat{\mu}$, the sample mean, and let $T_2' = \widehat{\sigma^2}$, the sample variance. Thus,

$$T_1' = \overline{X}_n \quad \text{and} \quad T_2' = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

We shall show that T_1' and T_2' are jointly sufficient statistics for μ and σ^2.

Let T_1 and T_2 be the jointly sufficient statistics for μ and σ^2 derived in Example 7.8.2. Then

$$T_1' = \frac{1}{n}T_1 \quad \text{and} \quad T_2' = \frac{1}{n}T_2 - \frac{1}{n^2}T_1^2.$$

Also, equivalently,

$$T_1 = nT_1' \quad \text{and} \quad T_2 = n(T_2' + T_1'^2).$$

Hence, the statistics T_1' and T_2' are obtained from the jointly sufficient statistics T_1 and T_2 by a one-to-one transformation. It follows, therefore, that T_1' and T_2' themselves are jointly sufficient statistics for μ and σ^2. ◀

We have now shown that the jointly sufficient statistics for the unknown mean and variance of a normal distribution can be chosen to be either T_1 and T_2, as given in Example 7.8.2, or T_1' and T_2', as given in Example 7.8.3.

Example 7.8.4 Jointly Sufficient Statistics for the Parameters of a Uniform Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[a, b]$, where the values of both endpoints a and b are unknown ($a < b$). The joint p.d.f. $f_n(\boldsymbol{x}|a, b)$ of $X_1, \ldots, X_n$ will be 0 unless all the observed values $x_1, \ldots, x_n$ lie between a and b; that is, $f_n(\boldsymbol{x}|a, b) = 0$ unless $\min\{x_1, \ldots, x_n\} \geq a$ and $\max\{x_1, \ldots, x_n\} \leq b$.

Furthermore, for every vector $\boldsymbol{x}$ such that $\min\{x_1, \ldots, x_n\} \ge a$ and $\max\{x_1, \ldots, x_n\} \le b$, we have

$$f_n(\boldsymbol{x}|a, b) = \frac{1}{(b-a)^n}.$$

For each two numbers y and z, we shall let $h(y, z)$ be defined as follows:

$$h(y, z) = \begin{cases} 1 & \text{for } y \le z, \\ 0 & \text{for } y > z. \end{cases}$$

For every value of $\boldsymbol{x} \in R^n$, we can then write

$$f_n(\boldsymbol{x}|a, b) = \frac{h[a, \min\{x_1, \ldots, x_n\}]\, h[\max\{x_1, \ldots, x_n\}, b]}{(b-a)^n}.$$

Since this expression depends on $\boldsymbol{x}$ only through the values of $\min\{x_1, \ldots, x_n\}$ and $\max\{x_1, \ldots, x_n\}$, it follows that the statistics $T_1 = \min\{X_1, \ldots, X_n\}$ and $T_2 = \max\{X_1, \ldots, X_n\}$ are jointly sufficient statistics for a and b. ◀

Minimal Sufficient Statistics

In a given problem, we want to try to find a sufficient statistic or a set of jointly sufficient statistics for θ, because the values of such statistics summarize all the relevant information about θ contained in the random sample. When a set of jointly sufficient statistics are known, the search for a good estimator of θ is simplified because we need consider only functions of these statistics as possible estimators. Therefore, in a given problem it is desirable to find, not merely any set of jointly sufficient statistics, but the *simplest* set of jointly sufficient statistics. That is, we want the set of sufficient statistics that requires us to consider the smallest collection of posible estimators. (We make this more precise in Defintion 7.8.3.) For example, it is correct but completely useless to say that in every problem the n observations $X_1, \ldots, X_n$ are jointly sufficient statistics.

We shall now describe another set of jointly sufficient statistics that exist in every problem and are slightly more useful.

Definition 7.8.2 Order Statistics. Suppose that $X_1, \ldots, X_n$ form a random sample from some distribution. Let Y_1 denote the smallest value in the random sample, let Y_2 denote the next smallest value, let Y_3 denote the third smallest value, and so on. In this way, Y_n denotes the largest value in the sample, and Y_{n-1} denotes the next largest value. The random variables $Y_1, \ldots, Y_n$ are called the *order statistics* of the sample.

Now let $y_1 \le y_2 \le \cdots \le y_n$ denote the values of the order statistics for a given sample. If we are told the values of $y_1, \ldots, y_n$, then we know that these n values were obtained in the sample. However, we do not know which one of the observations $X_1, \ldots, X_n$ actually yielded the value y_1, which one actually yielded the value y_2, and so on. All we know is that the smallest of the values of $X_1, \ldots, X_n$ was y_1, the next smallest value was y_2, and so on.

Theorem 7.8.2 Order Statistics Are Sufficient in Random Samples. Let $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$. Then the order statistics $Y_1, \ldots, Y_n$ are jointly sufficient for θ.

Proof Let $y_1 \le y_2 \le \cdots \le y_n$ denote the values of the order statistics. The joint p.d.f. or joint p.f. of $X_1, \ldots, X_n$ has the following form:

$$f_n(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{7.8.2}$$

Since the order of the factors in the product on the right side of Eq. (7.8.2) is irrelevant, Eq. (7.8.2) could just as well be rewritten in the form

$$f_n(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(y_i|\theta).$$

Hence, $f_n(\boldsymbol{x}|\theta)$ depends on $\boldsymbol{x}$ only through the values of $y_1, \ldots, y_n$. It follows, therefore, that the order statistics $Y_1, \ldots, Y_n$ are jointly sufficient statistics for θ. ■

In words, Theorem 7.8.2 says that it is sufficient to know the set of n numbers that were obtained in the sample, and it is not necessary to know which particular one of these numbers was, for example, the value of X_3.

To see how the order statistic is simpler than the full data vector in the sense of having fewer possible estimators, note that X_3 is an estimator based on the full data vector, but X_3 cannot be determined from the order statistics. Hence X_3 is not an estimator that we would need to consider if we based our inference on the order statistics. The same is true of all of the averages of the form $(X_{i_1} + \cdots + X_{i_k})/k$ for $\{i_1, \ldots, i_k\}$ a proper subset of $\{1, \ldots, n\}$, as well as many other functions. On the other hand, every estimator based on the order statistics is also a function of the full data.

In each of the examples that have been given in this section and in Sec. 7.7, we considered a distribution for which either there was a single sufficient statistic or there were two statistics that were jointly sufficient. For some distributions, however, the order statistics $Y_1, \ldots, Y_n$ are the simplest set of jointly sufficient statistics that exist, and no further reduction in terms of sufficient statistics is possible.

Example 7.8.5 Sufficient Statistics for the Parameter of a Cauchy Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from a Cauchy distribution centered at an unknown point θ $(-\infty < \theta < \infty)$. The p.d.f. $f(x|\theta)$ of this distribution is given by Eq. (7.6.6), and the joint p.d.f. $f_n(\boldsymbol{x}|\theta)$ of $X_1, \ldots, X_n$ is given by Eq. (7.6.7). It can be shown that the only jointly sufficient statistics that exist in this problem are the order statistics $Y_1, \ldots, Y_n$ or some other set of n statistics $T_1, \ldots, T_n$ that can be derived from the order statistics by a one-to-one transformation. The details of the argument will not be given here. ◂

These considerations lead us to the concepts of a minimal sufficient statistic and a minimal set of jointly sufficient statistics. A sufficient statistic T is a minimal sufficient statistic if every function of T, which itself is a sufficient statistic, is a one-to-one function of T. Formally, we shall use the following definition, which is equivalent to the informal definition just given.

Definition 7.8.3 Minimal (Jointly) Sufficient Statistic(s). A statistic T is a *minimal sufficient statistic* if T is sufficient and is a function of every other sufficient statistic. A vector $\boldsymbol{T} = (T_1, \ldots, T_k)$ of statistics are *minimal jointly sufficient statistics* if the coordinates of $\boldsymbol{T}$ are jointly sufficient statistics and $\boldsymbol{T}$ is a function of every other jointly sufficient statistics.

In Example 7.8.5, the order statistics $Y_1, \ldots, Y_n$ are minimal jointly sufficient statistics.

Maximum Likelihood Estimators and Bayes Estimators as Sufficient Statistics

For the next two theorems, let $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.f. or the p.d.f. is $f(x|\theta)$, where the value of the parameter θ is unknown and one-dimensional.

Theorem 7.8.3 M.L.E. and Sufficient Statistics. Let $T = r(X_1, \ldots, X_n)$ be a sufficient statistic for θ. Then the M.L.E. $\hat{\theta}$ of θ depends on the observations $X_1, \ldots, X_n$ only through the statistic T. Furthermore, if $\hat{\theta}$ is itself sufficient, then it is minimal sufficient.

Proof We show first that $\hat{\theta}$ is a function of every sufficient statistic. Let $T = r(\boldsymbol{X})$ be a sufficient statistic. The factorization criterion Theorem 7.7.1 says that the likelihood function $f_n(\boldsymbol{x}|\theta)$ can be written in the form

$$f_n(\boldsymbol{x}|\theta) = u(\boldsymbol{x})v[r(\boldsymbol{x}), \theta].$$

The M.L.E. $\hat{\theta}$ is the value of θ for which $f_n(\boldsymbol{x}|\theta)$ is a maximum. It follows, therefore, that $\hat{\theta}$ will be the value of θ for which $v[r(\boldsymbol{x}), \theta]$ is a maximum. Since $v[r(\boldsymbol{x}), \theta]$ depends on the observed vector $\boldsymbol{x}$ only through the function $r(\boldsymbol{x})$, it follows that $\hat{\theta}$ will also depend on $\boldsymbol{x}$ only through the function $r(\boldsymbol{x})$. Thus, the estimator $\hat{\theta}$ is a function of $T = r(\boldsymbol{X})$.

Since the estimator $\hat{\theta}$ is a function of the observations $X_1, \ldots, X_n$ and is not a function of the parameter θ, the estimator is itself a statistic. If $\hat{\theta}$ is actually a sufficient statistic, then it is minimal sufficient because we just showed that it is a function of every other sufficient statistic. ■

Theorem 7.8.3 can be extended easily to the case in which the parameter θ is multidimensional. If $\theta = (\theta_1, \ldots, \theta_k)$ is a vector of k real-valued parameters, then the M.L.E. vector $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ will depend on the observations $X_1, \ldots, X_n$ only through the functions in a set of jointly sufficient statistics. If the vecotor of the estimators $\hat{\theta}_1, \ldots, \hat{\theta}_k$ is a set of jointly sufficient statistics, then they are minimal jointly sufficient statistics because they are functions of every set of jointly sufficient statistics.

Example 7.8.6 Minimal Jointly Sufficient Statistics for the Parameters of a Normal Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which both the mean μ and the variance σ^2 are unknown. It was shown in Example 7.5.6 that the M.L.E.'s $\hat{\mu}$ and $\widehat{\sigma^2}$ are the sample mean and the sample variance. Also, it was shown in Example 7.8.3 that $\hat{\mu}$ and $\widehat{\sigma^2}$ are jointly sufficient statistics. Hence, $\hat{\mu}$ and $\widehat{\sigma^2}$ are minimal jointly sufficient statistics. ◄

The statistician in Example 7.8.6 can restrict the search for good estimators of μ and σ^2 to functions of minimal jointly sufficient statistics. It follows, therefore, from Example 7.8.6 that if the M.L.E.'s $\hat{\mu}$ and $\widehat{\sigma^2}$ themselves are not used as estimators of μ and σ^2, the only other estimators that need to be considered are functions of $\hat{\mu}$ and $\widehat{\sigma^2}$.

The results above concerning M.L.E.'s also pertain to Bayes estimators.

Theorem 7.8.4 Bayes Estimator and Sufficient Statistics. Let $T = r(\boldsymbol{X})$ be a sufficient statistic for θ. Then every Bayes estimator $\hat{\theta}$ of θ depends on the observations $X_1, \dots, X_n$ only through the statistic T. Furthermore, if $\hat{\theta}$ is itself sufficient, then it is minimal sufficient.

Proof Let the prior p.d.f. or p.f. of θ be $\xi(\theta)$. It follows from relation (7.2.10) and the factorization criterion that the posterior p.d.f. $\xi(\theta|x)$ will satisfy the following relation:

$$\xi(\theta|x) \propto v[r(\boldsymbol{x}), \theta]\xi(\theta).$$

It can be seen from this relation that the posterior p.d.f. of θ will depend on the observed vector $\boldsymbol{x}$ only through the value of $r(\boldsymbol{x})$. Since the Bayes estimator of θ with respect to a specified loss function is calculated from this posterior p.d.f., the estimator also will depend on the observed vector $\boldsymbol{x}$ only through the value of $r(\boldsymbol{x})$. In other words, the Bayes estimator is a function of $T = r(\boldsymbol{X})$. Since the Bayes estimator $\hat{\theta}$ is itself a statistic and is a function of every sufficient statistic T, if $\hat{\theta}$ is also sufficient, then it is minimal sufficient. ■

Theorem 7.8.4 also extends to vector parameters and jointly sufficient statistics.

Summary

Statistics $T_1 = r_1(\boldsymbol{X}), \dots, T_k = r_k(\boldsymbol{X})$ are jointly sufficient if and only if the joint p.f. or p.d.f. can be factored as $f_n(\boldsymbol{x}|\theta) = u(\boldsymbol{x})v[r_1(\boldsymbol{x}), \dots, r_k(\boldsymbol{x}), \theta]$, for some functions u and v. It is clear from this factorization that the original data $X_1, \dots, X_n$ are jointly sufficient. In order to be useful, a sufficient statistic should be a simpler function than the entire data. A minimal sufficient statistic is the simplest function that is still sufficient; that is, it is a sufficient statistic that is a function of every sufficient statistic. Since the likelihood function is a function of every sufficient statistic, according to the factorization criterion, a sufficient statistic that can be determined from the likelihood function is minimal sufficient. In particular, if an M.L.E. or Bayes estimator is sufficient, then it is minimal sufficient.

Exercises

Instructions for Exercises 1 to 4: In each exercise, assume that the random variables $X_1, \dots, X_n$ form a random sample of size n from the distribution specified in the exercise, and show that the statistics T_1 and T_2 specified in the exercise are jointly sufficient statistics.

1. A gamma distribution for which both parameters α and β are unknown ($\alpha > 0$ and $\beta > 0$); $T_1 = \prod_{i=1}^n X_i$ and $T_2 = \sum_{i=1}^n X_i$.

2. A beta distribution for which both parameters α and β are unknown ($\alpha > 0$ and $\beta > 0$); $T_1 = \prod_{i=1}^n X_i$ and $T_2 = \prod_{i=1}^n (1 - X_i)$.

3. A Pareto distribution (see Exercise 16 of Sec. 5.7) for which both parameters x_0 and α are unknown ($x_0 > 0$ and $\alpha > 0$); $T_1 = \min\{X_1, \dots, X_n\}$ and $T_2 = \prod_{i=1}^n X_i$.

4. The uniform distribution on the interval $[\theta, \theta + 3]$, where the value of θ is unknown ($-\infty < \theta < \infty$); $T_1 = \min\{X_1, \dots, X_n\}$ and $T_2 = \max\{X_1, \dots, X_n\}$.

5. Suppose that the vectors $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ form a random sample of two-dimensional vectors from a bivariate normal distribution for which the means, the variances, and the correlation are unknown. Show that the following five statistics are jointly sufficient: $\sum_{i=1}^n X_i$, $\sum_{i=1}^n Y_i$, $\sum_{i=1}^n X_i^2$, $\sum_{i=1}^n Y_i^2$, and $\sum_{i=1}^n X_i Y_i$.

6. Consider a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where the parameter θ is a k-dimensional vector belonging to some parameter space Ω. It is said that the family of distributions indexed by the values of θ in

Ω is a *k-parameter exponential family*, or a *k-parameter Koopman-Darmois family*, if $f(x|\theta)$ can be written as follows for $\theta \in \Omega$ and all values of x:

$$f(x|\theta) = a(\theta)b(x)\exp\left[\sum_{i=1}^{k} c_i(\theta)d_i(x)\right].$$

Here, a and $c_1, \dots, c_k$ are arbitrary functions of θ, and b and $d_1, \dots, d_k$ are arbitrary functions of x. Suppose now that $X_1, \dots, X_n$ form a random sample from a distribution which belongs to a k-parameter exponential family of this type, and define the k statistics $T_1, \dots, T_k$ as follows:

$$T_i = \sum_{j=1}^{n} d_i(X_j) \quad \text{for } i = 1, \dots, k.$$

Show that the statistics $T_1, \dots, T_k$ are jointly sufficient statistics for θ.

7. Show that each of the following families of distributions is a two-parameter exponential family as defined in Exercise 6:

a. The family of all normal distributions for which both the mean and the variance are unknown

b. The family of all gamma distributions for which both α and β are unknown

c. The family of all beta distributions for which both α and β are unknown

8. Suppose that $X_1, \dots, X_n$ form a random sample from an exponential distribution for which the value of the parameter β is unknown ($\beta > 0$). Is the M.L.E. of β a minimal sufficient statistic?

9. Suppose that $X_1, \dots, X_n$ form a random sample from the Bernoulli distribution with parameter p, which is unknown ($0 \le p \le 1$). Is the M.L.E. of p a minimal sufficient statistic?

10. Suppose that $X_1, \dots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown ($\theta > 0$). Is the M.L.E. of θ a minimal sufficient statistic?

11. Suppose that $X_1, \dots, X_n$ form a random sample from a Cauchy distribution centered at an unknown point θ ($-\infty < \theta < \infty$). Is the M.L.E. of θ a minimal sufficient statistic?

12. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution for which the p.d.f. is as follows:

$$f(x|\theta) = \begin{cases} \frac{2x}{\theta^2} & \text{for } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Here, the value of the parameter θ is unknown ($\theta > 0$). Determine the M.L.E. of the median of this distribution, and show that this estimator is a minimal sufficient statistic for θ.

13. Suppose that $X_1, \dots, X_n$ form a random sample from the uniform distribution on the interval $[a, b]$, where both endpoints a and b are unknown. Are the M.L.E.'s of a and b minimal jointly sufficient statistics?

14. For the conditions of Exercise 5, the M.L.E.'s of the means, the variances, and the correlation are given in Exercise 24 of Sec. 7.6. Are these five estimators minimal jointly sufficient statistics?

15. Suppose that $X_1, \dots, X_n$ form a random sample from the Bernoulli distribution with parameter p, which is unknown, and that the prior distribution of p is a certain specified beta distribution. Is the Bayes estimator of p with respect to the squared error loss function a minimal sufficient statistic?

16. Suppose that $X_1, \dots, X_n$ form a random sample from a Poisson distribution for which the value of the mean λ is unknown, and that the prior distribution of λ is a certain specified gamma distribution. Is the Bayes estimator of λ with respect to the squared error loss function a minimal sufficient statistic?

17. Suppose that $X_1, \dots, X_n$ form a random sample from a normal distribution for which the value of the mean μ is unknown and the value of the variance is known, and the prior distribution of μ is a certain specified normal distribution. Is the Bayes estimator of μ with respect to the squared error loss function a minimal sufficient statistic?

⋆ 7.9 Improving an Estimator

In this section, we show how to improve upon an estimator that is not a function of a sufficient statistic by using an estimator that is a function of a sufficient statistic.

The Mean Squared Error of an Estimator

Example 7.9.1

Customer Arrivals. A store owner is interested in the probability p that exactly one customer will arrive during a typical hour. She models customer arrivals as a Poisson process with rate θ per hour and observes how many customers arrive during each

of n hours, $X_1, \dots, X_n$. She converts each X_i to $Y_i = 1$ if $X_i = 1$ and $Y_i = 0$ if $X_i \neq 1$. Then $Y_1, \dots, Y_n$ is a random sample from the Bernoulli distribution with parameter p. The store owner then estimates p by $\delta(\boldsymbol{X}) = \sum_{i=1}^{n} Y_i/n$. Is this a good estimator? In particular, if the store owner wants to minimize mean squared error, is there another estimator that we can show is better? ◀

In general, suppose that $\boldsymbol{X} = (X_1, \dots, X_n)$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where the parameter θ must belong to some parameter space Ω. In this section, θ can be a one-dimensional parameter or a vector of parameters. For each random variable $Z = g(X_1, \dots, X_n)$, we shall let $E_\theta(Z)$ denote the expectation of Z calculated with respect to the joint p.d.f. or joint p.f. $f_n(\boldsymbol{x}|\theta)$. If we were thinking of θ as a random variable, then $E_\theta(Z) = E(Z|\theta)$. For example, if $f_n(\boldsymbol{x}|\theta)$ is a p.d.f., then

$$E_\theta(Z) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\boldsymbol{x}) f_n(\boldsymbol{x}|\theta)\, dx_1 \dots dx_n.$$

We shall suppose that the value of θ is unknown and that we want to estimate some function $h(\theta)$. If θ is a vector, $h(\theta)$ might be one of the coordinates or a function of all coordinates, and so on. We shall assume that the squared error loss function is to be used. Also, for each given estimator $\delta(\boldsymbol{X})$ and every given value of $\theta \in \Omega$, we shall let $R(\theta, \delta)$ denote the M.S.E. of δ calculated with respect to the given value of θ. Thus,

$$R(\theta, \delta) = E_\theta([\delta(\boldsymbol{X}) - h(\theta)]^2). \tag{7.9.1}$$

If we do not assign a prior distribution to θ, then it is desired to find an estimator δ for which the M.S.E. $R(\theta, \delta)$ is small for every value of $\theta \in \Omega$ or, at least, for a wide range of values of θ.

Suppose now that $\boldsymbol{T}$ is a vector of jointly sufficient statistics for θ. In the remainder of this section we shall refer to $\boldsymbol{T}$ simply as the sufficient statistic. If $\boldsymbol{T}$ is one-dimensional, just pretend that we wrote it as T. Consider a statistician A who plans to use a particular estimator $\delta(\boldsymbol{X})$. In Sec. 7.7 we remarked that another statistician B who learns only the value of the sufficient statistic $\boldsymbol{T}$ can generate, by means of an auxiliary randomization, an estimator that will have exactly the same distribution as $\delta(\boldsymbol{X})$ and, in particular, will have the same mean squared error as $\delta(\boldsymbol{X})$ for every value of $\theta \in \Omega$. We shall now show that even without using an auxiliary randomization, statistician B can find an estimator δ_0 that depends on the observations $\boldsymbol{X}$ only through the sufficient statistic $\boldsymbol{T}$ and is at least as good an estimator as δ in the sense that $R(\theta, \delta_0) \leq R(\theta, \delta)$, for every value of $\theta \in \Omega$.

Conditional Expectation When a Sufficient Statistic Is Known

We shall define the estimator $\delta_0(\boldsymbol{T})$ by the following conditional expectation:

$$\delta_0(\boldsymbol{T}) = E_\theta[\delta(\boldsymbol{X})|\boldsymbol{T}]. \tag{7.9.2}$$

Since $\boldsymbol{T}$ is a sufficient statistic, the conditional joint distribution of $X_1, \dots, X_n$ for each given value of $\boldsymbol{T}$ is the same for every value of $\theta \in \Omega$. Therefore, for any given value of $\boldsymbol{T}$, the conditional expectation of the function $\delta(\boldsymbol{X})$ will be the same for every value of $\theta \in \Omega$. It follows that the conditional expectation in Eq. (7.9.2) will depend on the value of $\boldsymbol{T}$ but will not actually depend on the value of θ. In other words, the function $\delta_0(\boldsymbol{T})$ is indeed an estimator of θ because it depends only on the observations $\boldsymbol{X}$ and does not depend on the unknown value of θ. For this reason, we

can omit the subscript θ on the expectation symbol E in Eq. (7.9.2), and we can write the relation as follows:

$$\delta_0(\boldsymbol{T}) = E[\delta(\boldsymbol{X})|\boldsymbol{T}]. \tag{7.9.3}$$

We can now prove the following theorem, which was established independently by D. Blackwell and C. R. Rao in the late 1940s.

Theorem 7.9.1 Let $\delta(\boldsymbol{X})$ be an estimator, let $\boldsymbol{T}$ be a sufficient statistic for θ, and let the estimator $\delta_0(\boldsymbol{T})$ be defined as in Eq. (7.9.3). Then for every value of $\theta \in \Omega$,

$$R(\theta, \delta_0) \le R(\theta, \delta). \tag{7.9.4}$$

Furthermore, if $R(\theta, \delta) < \infty$, there is strict inequality in (7.9.4) unless $\delta(\boldsymbol{X})$ is a function of $\boldsymbol{T}$.

Proof If the M.S.E. $R(\theta, \delta)$ is infinite for a given value of $\theta \in \Omega$, then the relation (7.9.4) is automatically satisfied. We shall assume, therefore, that $R(\theta, \delta) < \infty$. It follows from part (a) of Exercise 4 in Sec. 4.4 that

$$E_\theta([\delta(\boldsymbol{X}) - \theta]^2) \ge (E_\theta[\delta(\boldsymbol{X})] - \theta)^2,$$

and it can be shown that this same relationship must also hold if the expectations are replaced by conditional expectations given $\boldsymbol{T}$. Therefore,

$$E_\theta([\delta(\boldsymbol{X}) - \theta]^2|\boldsymbol{T}) \ge (E_\theta[\delta(\boldsymbol{X})|\boldsymbol{T}] - \theta)^2 = [\delta_0(\boldsymbol{T}) - \theta]^2. \tag{7.9.5}$$

It now follows from relation (7.9.5) that

$$\begin{aligned} R(\theta, \delta_0) &= E_\theta[\{\delta_0(\boldsymbol{T}) - \theta\}^2] \le E_\theta\{E_\theta[\{\delta(\boldsymbol{X}) - \theta\}^2|\boldsymbol{T}]\} \\ &= E_\theta[\{\delta(\boldsymbol{X}) - \theta\}^2] = R(\theta, \delta), \end{aligned}$$

where the next-to-last equality follows from Theorem 4.7.1, the law of total probability for expectations. Hence, $R(\theta, \delta_0) \le R(\theta, \delta)$ for every value of $\theta \in \Omega$.

Finally, suppose that $R(\theta, \delta) < \infty$ and that $\delta(\boldsymbol{X})$ is not a function of $\boldsymbol{T}$. That is, there is no function $g(\boldsymbol{T})$ such that $\Pr(\delta(\boldsymbol{X}) = g(\boldsymbol{T})|\boldsymbol{T}) = 1$. Then part (b) of Exercise 4 in Sec. 4.4 (conditional on $\boldsymbol{T}$) says that there is strict inequality in (7.9.4). ■

Example 7.9.2 Customer Arrivals. Return now to Example 7.9.1. Let θ stand for the rate of customer arrivals in units per hour. Then $\boldsymbol{X}$ forms a random sample from the Poisson distribution with mean θ. Example 7.7.2 shows us that a sufficient statistic is $T = \sum_{i=1}^n X_i$. The distribution of T is the Poisson distribution with mean $n\theta$. We shall now compute

$$\delta_0(T) = E[\delta(\boldsymbol{X})|T],$$

where $\delta(\boldsymbol{X}) = \sum_{i=1}^n Y_i/n$ was defined in Example 7.9.1. (Recall that $Y_i = 1$ if $X_i = 1$ and $Y_i = 0$ if $X_i \ne 1$ so that $\delta(\boldsymbol{X})$ is the proportion of hours in which exactly one customer arrives.) For each i and each possible value t of T, it is easy to see that

$$E(Y_i|T = t) = \Pr(X_i = 1|T = t) = \frac{\Pr(X_i = 1, T = t)}{\Pr(T = t)} = \frac{\Pr\left(X_i = 1, \sum_{j \ne i} X_j = t - 1\right)}{\Pr(T = t)}.$$

For $t = 0$, $\Pr(X_i = 1|T = 0) = 0$ trivially. For $t > 0$, we see that

$$\Pr(T = t) = \frac{e^{-n\theta}(n\theta)^t}{t!},$$

$$\Pr\left(X_i = 1, \sum_{j \neq i} X_j = t - 1\right) = e^{-\theta}\theta \times \frac{e^{-[n-1]\theta}([n-1]\theta)^{t-1}}{(t-1)!} = \frac{e^{-n\theta}[n-1]^{t-1}\theta^t}{(t-1)!}.$$

The ratio of these two probabilities is

$$E(Y_i|T = t) = \frac{t}{n}\left(1 - \frac{1}{n}\right)^{t-1}. \tag{7.9.6}$$

It follows that

$$\delta_0(t) = E[\delta_0(\boldsymbol{x})|T = t] = E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i \,\middle|\, T = t\right] = \frac{1}{n}\sum_{i=1}^{n} E(Y_i|T = t).$$

According to Eq. (7.9.6), all $E(Y_i|T = t)$ are the same, so $\delta_0(t)$ is the right-hand side of Eq. (7.9.6). That $\delta_0(T)$ is better than $\delta(\boldsymbol{X})$ under squared error loss follows from Theorem 7.9.1. ◀

A result similar to Theorem 7.9.1 holds if $R(\theta, \delta)$ is defined as the M.A.E. of an estimator for a given value of $\theta \in \Omega$ instead of the M.S.E. of δ. In other words, suppose that $R(\theta, \delta)$ is defined as follows:

$$R(\theta, \delta) = E_\theta(|\delta(\boldsymbol{X}) - \theta|). \tag{7.9.7}$$

Then it can be shown (see Exercise 10 at the end of this section) that Theorem 7.9.1 is still true.

Definition 7.9.1 Inadmissible/Admissible/Dominates. Suppose that $R(\theta, \delta)$ is defined by either Eq. (7.9.1) or Eq. (7.9.7). It is said that an estimator δ is *inadmissible* if there exists another estimator δ_0 such that $R(\theta, \delta_0) \le R(\theta, \delta)$ for every value of $\theta \in \Omega$ and there is strict inequality in this relation for at least one value of $\theta \in \Omega$. Under these conditions, it is also said that the estimator δ_0 *dominates* the estimator δ. An estimator δ_0 is *admissible* if there is no other estimator that dominates δ_0.

In the terminology of Definition 7.9.1, Theorem 7.9.1 can be summarized as follows: An estimator δ that is not a function of the sufficient statistic $\boldsymbol{T}$ alone must be inadmissible. Theorem 7.9.1 also explicitly identifies an estimator $\delta_0 = E(\delta(\boldsymbol{X})|\boldsymbol{T})$ that dominates δ. However, this part of the theorem is somewhat less useful in a practical problem, because it is usually very difficult to calculate the conditional expectation $E(\delta(\boldsymbol{X})|\boldsymbol{T})$. Theorem 7.9.1 is valuable mainly because it provides further strong evidence that we can restrict our search for a good estimator of θ to those estimators that depend on the observations only through a sufficient statistic.

Example 7.9.3 Estimating the Mean of a Normal Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which the mean μ is unknown and the variance is known, and let $Y_1 \le \cdots \le Y_n$ denote the order statistics of the sample, as defined in Sec. 7.8. If n is an odd number, then the middle observation $Y_{(n+1)/2}$ is called the *sample median*. If n is an even number, then each value between the two middle observations $Y_{n/2}$ and $Y_{(n/2)+1}$ is a *sample median*, but the particular value $\frac{1}{2}[Y_{n/2} + Y_{(n/2)+1}]$ is often referred to as *the sample median*.

Since the normal distribution from which the sample is drawn is symmetric with respect to the point μ, the median of the normal distribution is μ. Therefore, we might consider the use of the sample median, or a simple function of the sample median, as an estimator of μ. However, it was shown in Example 7.7.4 that the sample mean $\overline{X}_n$ is a sufficient statistic for μ. It follows from Theorem 7.9.1 that every function of the sample median that might be used as an estimator of μ will be dominated by some other function of $\overline{X}_n$. In searching for an estimator of μ, we need consider only functions of $\overline{X}_n$. ◀

Example 7.9.4

Estimating the Standard Deviation of a Normal Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which both the mean μ and the variance σ^2 are unknown, and again let $Y_1 \le \cdots \le Y_n$ denote the order statistics of the sample. The difference $Y_n - Y_1$ is called the *range* of the sample, and we might consider using some simple function of the range as an estimator of the standard deviation σ. However, it was shown in Example 7.8.2 that the statistics $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i^2$ are jointly sufficient for the parameters μ and σ^2. Therefore, every function of the range that might be used as an estimator of σ will be dominated by a function of $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i^2$. ◀

Example 7.9.5

Failure Times of Ball Bearings. Suppose that we wish to estimate the mean failure time of the ball bearings described in Example 5.6.9 based on the sample of 23 observed failure times. Let $Y_1, \ldots, Y_{23}$ be the observed failure times (not the logarithms). We might consider using the average $\overline{Y}_n = \frac{1}{23}\sum_{i=1}^{23} Y_i$ as an estimator. Suppose that we continue to model the logarithms $X_i = \log(Y_i)$ as normal random variables with mean θ and variance 0.25. Then Y_i has the lognormal distribution with parameters θ and 0.25. From Eq. (5.6.15), the mean of Y_i is $\exp(\theta + 0.125)$, the mean failure time. However, we know that $\overline{X}_n$ is sufficient. Since $\overline{Y}_n$ is not a function of $\overline{X}_n$, there is a function of $\overline{X}_n$ that improves on $\overline{Y}_n$ as an estimator of the mean failure time. We can actually find which function that is. First, write

$$E(\overline{Y}_n|\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i|\overline{X}_n). \tag{7.9.8}$$

In Exercise 15 of Sec. 5.10, you proved that the conditional distribution of X_i given $\overline{X}_n = \overline{x}_n$ is the normal distribution with mean $\overline{x}_n$ and variance $0.25(1 - 1/n)$ for every i. It follows that, for each i, the conditional distribution of Y_i given $\overline{X}_n$ is the lognormal distribution with parameters $\overline{X}_n$ and $0.25(1 - 1/n)$. Hence, it follows from Eq. (5.6.15) that the conditional mean of Y_i given $\overline{X}_n$ is $\exp[\overline{X}_n + 0.125(1 - 1/n)]$ for all i, and Eq. (7.9.8) equals $\exp[\overline{X}_n + 0.125(1 - 1/n)]$ as well. ◀

Limitation of the Use of Sufficient Statistics

When the foregoing theory of sufficient statistics is applied in a statistical problem, it is important to keep in mind the following limitation. The existence and the form of a sufficient statistic in a particular problem depend critically on the form of the function assumed for the p.d.f. or the p.f. A statistic that is a sufficient statistic when it is assumed that the p.d.f. is $f(x|\theta)$ may not be a sufficient statistic when it is assumed that the p.d.f. is $g(x|\theta)$, even though $g(x|\theta)$ may be quite similar to $f(x|\theta)$ for every value of $\theta \in \Omega$. Suppose that a statistician is in doubt about the exact form of the p.d.f. in a specific problem but assumes for convenience that the p.d.f. is $f(x|\theta)$; suppose also that the statistic $\boldsymbol{T}$ is a sufficient statistic under this assumption. Because of the

statistician's uncertainty about the exact form of the p.d.f., he may wish to use an estimator of θ that performs reasonably well for a wide variety of possible p.d.f.'s, even though the selected estimator may not meet the requirement that it should depend on the observations only through the statistic $\boldsymbol{T}$.

An estimator that performs reasonably well for a wide variety of possible p.d.f.'s, even though it may not necessarily be the best available estimator for any particular family of p.d.f.'s, is often called a *robust estimator*. We shall consider robust estimators further in Chapter 10.

The preceding discussion also raises another useful point to keep in mind. In Sec. 7.2, we introduced *sensitivity analysis* as a way to study the effect of the choice of prior distribution on an inference. The same idea can be applied to any feature of a statistical model that is chosen by a statistician. In particular, the distribution for the observations given the parameters, as defined through $f(x|\theta)$, is often chosen for convenience rather than through a careful analysis. One can perform an inference repeatedly using different distributions for the observable data. The comparison of the resulting inferences from each choice is another form of sensitivity analysis.

Summary

Suppose that $\boldsymbol{T}$ is a sufficient statistic, and we are trying to estimate a parameter with squared error loss. Suppose that an estimator $\delta(\boldsymbol{X})$ is not a function of $\boldsymbol{T}$. Then δ can be improved by using $\delta_0(\boldsymbol{T})$, the conditional mean of $\delta(\boldsymbol{X})$ given $\boldsymbol{T}$. Because $\delta_0(\boldsymbol{T})$ has the same mean as $\delta(\boldsymbol{X})$ and its variance is no larger, it follows that $\delta_0(\boldsymbol{T})$ has M.S.E. that is no larger than that of $\delta(\boldsymbol{X})$.

Exercises

1. Suppose that the random variables $X_1, \dots, X_n$ form a random sample of size n ($n \ge 2$) from the normal distribution with mean 0 and unknown variance θ. Suppose also that for every estimator $\delta(X_1, \dots, X_n)$, the M.S.E. $R(\theta, \delta)$ is defined by Eq. (7.9.1). Explain why the sample variance is an inadmissible estimator of θ.

2. Suppose that the random variables $X_1, \dots, X_n$ form a random sample of size n ($n \ge 2$) from the uniform distribution on the interval $[0, \theta]$, where the value of the parameter θ is unknown ($\theta > 0$) and must be estimated. Suppose also that for every estimator $\delta(X_1, \dots, X_n)$, the M.S.E. $R(\theta, \delta)$ is defined by Eq. (7.9.1). Explain why the estimator $\delta_1(X_1, \dots, X_n) = 2\overline{X}_n$ is inadmissible.

3. Consider again the conditions of Exercise 2, and let the estimator δ_1 be as defined in that exercise. Determine the value of the M.S.E. $R(\theta, \delta_1)$ for $\theta > 0$.

4. Consider again the conditions of Exercise 2. Let $Y_n = \max\{X_1, \dots, X_n\}$ and consider the estimator $\delta_2(X_1, \dots, X_n) = Y_n$.

a. Determine the M.S.E. $R(\theta, \delta_2)$ for $\theta > 0$.

b. Show that for $n = 2$, $R(\theta, \delta_2) = R(\theta, \delta_1)$ for $\theta > 0$.

c. Show that for $n \ge 3$, the estimator δ_2 dominates the estimator δ_1.

5. Consider again the conditions of Exercises 2 and 4. Show that there exists a constant c^* such that the estimator $c^* Y_n$ dominates every other estimator having the form cY_n for $c \ne c^*$.

6. Suppose that $X_1, \dots, X_n$ form a random sample of size n ($n \ge 2$) from the gamma distribution with parameters α and β, where the value of α is unknown ($\alpha > 0$) and the value of β is known. Explain why $\overline{X}_n$ is an inadmissible estimator of the mean of this distribution when the squared error loss function is used.

7. Suppose that $X_1, \dots, X_n$ form a random sample from an exponential distribution for which the value of the parameter β is unknown ($\beta > 0$) and must be estimated by using the squared error loss function. Let δ be the estimator such that $\delta(X_1, \dots, X_n) = 3$ for all possible values of $X_1, \dots, X_n$.

a. Determine the value of the M.S.E. $R(\beta, \delta)$ for $\beta > 0$.

b. Explain why the estimator δ must be admissible.

8. Suppose that a random sample of n observations is taken from a Poisson distribution for which the value of the mean θ is unknown ($\theta > 0$), and the value of $\beta = e^{-\theta}$ must be estimated by using the squared error loss function. Since β is equal to the probability that an observation from this Poisson distribution will have the value 0, a natural estimator of β is the proportion $\hat{\beta}$ of observations in the random sample that have the value 0. Explain why $\hat{\beta}$ is an inadmissible estimator of β.

9. For every random variable X, show that $|E(X)| \le E(|X|)$.

10. Let $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where $\theta \in \Omega$. Suppose that the value of θ must be estimated, and that T is a sufficient statistic for θ. Let δ be an arbitrary estimator of θ, and let δ_0 be another estimator defined by the relation $\delta_0 = E(\delta|T)$. Show that for every value of $\theta \in \Omega$,

$$E_\theta(|\delta_0 - \theta|) \le E_\theta(|\delta - \theta|).$$

11. Suppose that the variables $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where $\theta \in \Omega$, and let $\hat{\theta}$ denote the M.L.E. of θ. Suppose also that the statistic T is a sufficient statistic for θ, and let the estimator δ_0 be defined by the relation $\delta_0 = E(\hat{\theta}|T)$. Compare the estimators $\hat{\theta}$ and δ_0.

12. Suppose that $X_1, \ldots, X_n$ form a sequence of n Bernoulli trials for which the probability p of success on any given trial is unknown ($0 \le p \le 1$), and let $T = \sum_{i=1}^n X_i$. Determine the form of the estimator $E(X_1|T)$.

13. Suppose that $X_1, \ldots, X_n$ form a random sample from a Poisson distribution for which the value of the mean θ is unknown ($\theta > 0$). Let $T = \sum_{i=1}^n X_i$, and for $i = 1, \ldots, n$, let the statistic Y_i be defined as follows:

$$Y_i = \begin{cases} 1 & \text{if } X_i = 0, \\ 0 & \text{if } X_i > 0. \end{cases}$$

Determine the form of the estimator $E(Y_i|T)$.

14. Consider again the conditions of Exercise 8. Determine the form of the estimator $E(\hat{\beta}|T)$. You may wish to use results obtained while solving Exercise 13.

15. Find the M.L.E. of $\exp(\theta + 0.125)$ in Example 7.9.5. Both the M.L.E. and the estimator in Example 7.9.5 have the form $\exp(\overline{X}_n + c)$ for some constant c. Find the value c so that the estimator $\exp(\overline{X}_n + c)$ has the smallest possible M.S.E.

16. In Example 7.9.1, find the formula for p in terms of θ, the mean of each X_i. Also find the M.L.E. of p and show that the estimator $\delta_0(T)$ in Example 7.9.2 is nearly the same as the M.L.E. if n is large.

7.10 Supplementary Exercises

1. A program will be run with 25 different sets of input. Let θ stand for the probability that an execution error will occur during a single run. We believe that, conditional on θ, each run of the program will encounter an error with probability θ and that the different runs are independent. Prior to running the program, we believe that θ has the uniform distribution on the interval [0, 1]. Suppose that we get errors during 10 of the 25 runs.

- **a.** Find the posterior distribution of θ.
- **b.** If we wanted to estimate θ by $\hat{\theta}$ using squared error loss, what would our estimate $\hat{\theta}$ be?

2. Suppose that $X_1, \ldots, X_n$ are i.i.d. with $\Pr(X_i = 1) = \theta$ and $\Pr(X_i = 0) = 1 - \theta$, where θ is unknown ($0 \le \theta \le 1$). Find the M.L.E. of θ^2.

3. Suppose that the proportion θ of bad apples in a large lot is unknown and has the following prior p.d.f.:

$$\xi(\theta) = \begin{cases} 60\theta^2(1-\theta)^3 & \text{for } 0 < \theta < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that a random sample of 10 apples is drawn from the lot, and it is found that three are bad. Find the Bayes estimate of θ with respect to the squared error loss function.

4. Suppose that $X_1, \ldots, X_n$ form a random sample from a uniform distribution with the following p.d.f.:

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{for } \theta \le x \le 2\theta, \\ 0 & \text{otherwise.} \end{cases}$$

Assuming that the value of θ is unknown ($\theta > 0$), determine the M.L.E. of θ.

5. Suppose that X_1 and X_2 are independent random variables, and that X_i has the normal distribution with mean $b_i\mu$ and variance σ_i^2 for $i = 1, 2$. Suppose also that b_1, b_2, σ_1^2, and σ_2^2 are known positive constants, and that μ is an unknown parameter. Determine the M.L.E. of μ based on X_1 and X_2.

6. Let $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ for $\alpha > 0$ (the digamma function). Show that

$$\psi(\alpha + 1) = \psi(\alpha) + \frac{1}{\alpha}.$$

7. Suppose that a regular light bulb, a long-life light bulb, and an extra-long-life light bulb are being tested. The lifetime X_1 of the regular bulb has the exponential distribution with mean θ, the lifetime X_2 of the long-life bulb has the exponential distribution with mean 2θ, and the lifetime X_3 of the extra-long-life bulb has the exponential distribution with mean 3θ.

a. Determine the M.L.E. of θ based on the observations X_1, X_2, and X_3.

b. Let $\psi = 1/\theta$, and suppose that the prior distribution of ψ is the gamma distribution with parameters α and β. Determine the posterior distribution of ψ given X_1, X_2, and X_3.

8. Consider a Markov chain with two possible states s_1 and s_2 and with stationary transition probabilities as given in the following transition matrix $\boldsymbol{P}$:

$$\boldsymbol{P} = \begin{array}{c} \\ s_1 \\ s_2 \end{array} \begin{array}{c} \begin{array}{cc} s_1 & s_2 \end{array} \\ \begin{bmatrix} \theta & 1-\theta \\ 3/4 & 1/4 \end{bmatrix} \end{array},$$

where the value of θ is unknown ($0 \le \theta \le 1$). Suppose that the initial state X_1 of the chain is s_1, and let $X_2, \dots, X_{n+1}$ denote the state of the chain at each of the next n successive periods. Determine the M.L.E. of θ based on the observations $X_2, \dots, X_{n+1}$.

9. Suppose that an observation X is drawn from a distribution with the following p.d.f.:

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 < x < \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Also, suppose that the prior p.d.f. of θ is

$$\xi(\theta) = \begin{cases} \theta e^{-\theta} & \text{for } \theta > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Determine the Bayes estimator of θ with respect to **(a)** the mean squared error loss function and **(b)** the absolute error loss function.

10. Suppose that $X_1, \dots, X_n$ form n Bernoulli trials with parameter $\theta = (1/3)(1 + \beta)$, where the value of β is unknown ($0 \le \beta \le 1$). Determine the M.L.E. of β.

11. The method of *randomized response* is sometimes used to conduct surveys on sensitive topics. A simple version of the method can be described as follows: A random sample of n persons is drawn from a large population. For each person in the sample there is probability 1/2 that the person will be asked a standard question and probability 1/2 that the person will be asked a sensitive question. Furthermore, this selection of the standard or the sensitive question is made independently from person to person. If a person is asked the standard question, then there is probability 1/2 that she will give a positive response; however if she is asked the sensitive question, then there is an unknown probability p that she will give a positive response. The statistician can observe only the total number X of positive responses that were given by the n persons in the sample. He cannot observe which of these persons were asked the sensitive question or how many persons in the sample were asked the sensitive question. Determine the M.L.E. of p based on the observation X.

12. Suppose that a random sample of four observations is to be drawn from the uniform distribution on the interval $[0, \theta]$, and that the prior distribution of θ has the following p.d.f.:

$$\xi(\theta) = \begin{cases} \frac{1}{\theta^2} & \text{for } \theta \ge 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that the values of the observations in the sample are found to be 0.6, 0.4, 0.8, and 0.9. Determine the Bayes estimate of θ with respect to the squared error loss function.

13. For the conditions of Exercise 12, determine the Bayes estimate of θ with respect to the absolute error loss function.

14. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution with the following p.d.f.:

$$f(x|\beta, \theta) = \begin{cases} \beta e^{-\beta(x-\theta)} & \text{for } x \ge \theta, \\ 0 & \text{otherwise,} \end{cases}$$

where β and θ are unknown ($\beta > 0$, $-\infty < \theta < \infty$). Determine a pair of jointly sufficient statistics.

15. Suppose that $X_1, \dots, X_n$ form a random sample from the Pareto distribution with parameters x_0 and α (see Exercise 16 of Sec. 5.7), where x_0 is unknown and α is known. Determine the M.L.E. of x_0.

16. Determine whether the estimator found in Exercise 15 is a minimal sufficient statistic.

17. Consider again the conditions of Exercise 15, but suppose now that both parameters x_0 and α are unknown. Determine the M.L.E.'s of x_0 and α.

18. Determine whether the estimators found in Exercise 17 are minimal jointly sufficient statistics.

19. Suppose that the random variable X has a binomial distribution with an unknown value of n and a known value of p ($0 < p < 1$). Determine the M.L.E. of n based on the observation X. *Hint:* Consider the ratio

$$\frac{f(x|n+1, p)}{f(x|n, p)}.$$

20. Suppose that two observations X_1 and X_2 are drawn at random from a uniform distribution with the following p.d.f.:

$$f(x|\theta) = \begin{cases} \frac{1}{2\theta} & \text{for } 0 \le x \le \theta \text{ or } 2\theta \le x \le 3\theta, \\ 0 & \text{otherwise,} \end{cases}$$

where the value of θ is unknown ($\theta > 0$). Determine the M.L.E. of θ for each of the following pairs of observed values of X_1 and X_2:

a. $X_1 = 7$ and $X_2 = 9$

b. $X_1 = 4$ and $X_2 = 9$

c. $X_1 = 5$ and $X_2 = 9$

21. Suppose that a random sample $X_1, \ldots, X_n$ is to be taken from the normal distribution with unknown mean θ and variance 100, and the prior distribution of θ is the normal distribution with specified mean μ_0 and variance 25. Suppose that θ is to be estimated using the squared error loss function, and the sampling cost of each observation is 0.25 (in appropriate units). If the total cost of the estimation procedure is equal to the expected loss of the Bayes estimator plus the sampling cost $(0.25)n$, what is the sample size n for which the total cost will be a minimum?

22. Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with unknown mean θ, and the variance of this distribution is to be estimated using the squared error loss function. Determine whether or not the sample variance is an admissible estimator.

23. The formulas (7.5.6) for the sample mean and sample variance are of theoretical importance, but they can be inefficient or produce inaccurate results if used for numerical calculation with very large samples. For example, let $x_1, x_2, \ldots$ be a sequence of real numbers. Computing $\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$ directly requires that we first compute $\overline{x}_n$ and then still have all n observations available so that we can compute $x_i - \overline{x}_n$ for each i. Also, if n is very large, then computing $\overline{x}_n$ by adding the x_i's together can produce large rounding errors once the next x_i becomes very small relative to the accummulated sum.

a. Prove the seemingly more efficient formula

$$\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}_n^2.$$

With this formula, we could accummulate the sum of the x_i's and x_i^2's separately and forget each observation afterward. We would still suffer the rounding error problem mentioned above.

b. Prove the following formulas that reduce the rounding error problem in accummulating a sum. For each integer n

$$\overline{x}_{n+1} = \overline{x}_n + \frac{1}{n+1}(x_{n+1} - \overline{x}_n),$$

$$\sum_{i=1}^{n+1}(x_i - \overline{x}_{n+1})^2 = \sum_{i=1}^{n}(x_i - \overline{x}_n)^2 + \frac{n}{n+1}(x_{n+1} - \overline{x}_n)^2.$$

These formulas allow us to forget each x_i after we use it to update the two formulas.

Sampling Distributions of Estimators

Chapter 8

8.1 The Sampling Distribution of a Statistic

A statistic is a function of some observable random variables, and hence is itself a random variable with a distribution. That distribution is its sampling distribution, and it tells us what values the statistic is likely to assume and how likely it is to assume those values prior to observing our data. When the distribution of the observable data is indexed by a parameter, the sampling distribution is specified as the distribution of the statistic for a given value of the parameter.

Statistics and Estimators

Example 8.1.1 A Clinical Trial. In the clinical trial first introduced in Example 2.1.4, let θ stand for the proportion who do not relapse among all possible imipramine patients. We could use the observed proportion of patients without relapse in the imipramine group to estimate θ. Prior to observing the data, the proportion of sampled patients with no relapse is a random variable T that has a distribution and will not exactly equal the parameter θ. However, we hope that T will be close to θ with high probability. For example, we could try to compute the probability that $|T - \theta| < 0.1$. Such calculations require that we know the distribution of the random variable T. In the clinical trial, we modeled the responses of the 40 patients in the imipramine group as conditionally (given θ) i.i.d. Bernoulli random variables with parameter θ. It follows that the conditional distribution of $40T$ given θ is the binomial distribution with parameters 40 and θ. The distribution of T can be derived easily from this. Indeed T has the following p.f. given θ:

$$f(t|\theta) = \binom{40}{40t}\theta^{40t}(1-\theta)^{40(1-t)}, \quad \text{for } t = 0, \tfrac{1}{40}, \tfrac{2}{40}, \ldots, \tfrac{39}{40}, 1,$$

and $f(t|\theta) = 0$ otherwise. ◄

The distribution at the end of Example 8.1.1 is called the *sampling distribution* of the statistic T, and we can use it to help address questions such as how close we expect T to be to θ prior to observing the data. We can also use the sampling distribution of T to help to determine how much we will learn about θ by observing T. If we are

trying to decide which of two different statistics to use as an estimator, their sampling distributions can be useful for helping us to compare them.

The concept of sampling distribution applies to a larger class of random variables than statistics.

Definition 8.1.1 Sampling Distribution. Suppose that the random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from a distribution involving a parameter θ whose value is unknown. Let T be a function of $\boldsymbol{X}$ and possibly θ. That is, $T = r(X_1, \ldots, X_n, \theta)$. The distribution of T (given θ) is called the *sampling distribution* of T. We will use the notation $E_\theta(T)$ to denote the mean of T calculated from its sampling distribution.

The name "sampling distribution" comes from the fact that T depends on a random sample and so its distribution is derived from the distribution of the sample.

Often, the random variable T in Definition 8.1.1 will not depend on θ, and hence will be a statistic as defined in Definition 7.1.4. In particular, if T is an estimator of θ (as defined in Definition 7.4.1), then T is also a statistic because it is a function of $\boldsymbol{X}$. Therefore, in principle, it is possible to derive the sampling distribution of each estimator of θ. In fact, the distributions of many estimators and statistics have already been found in previous parts of this book.

Example 8.1.2 Sampling Distribution of the M.L.E. of the Mean of a Normal Distribution. Supppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. We found in Examples 7.5.5 and 7.5.6 that the sample mean $\overline{X}_n$ is the M.L.E. of μ. Furthermore, it was found in Corollary 5.6.2 that the distribution of $\overline{X}_n$ is the normal distribution with mean μ and variance σ^2/n. ◀

In this chapter, we shall derive, for random samples from a normal distribution, the distribution of the sample variance and the distributions of various functions of the sample mean and the sample variance. These derivations will lead us to the definitions of some new distributions that play important roles in problems of statistical inference. In addition, we shall study certain general properties of estimators and their sampling distributions.

Purpose of the Sampling Distribution

Example 8.1.3 Lifetimes of Electronic Components. Consider the company in Example 7.1.1 that sells electronic components. They model the lifetimes of these components as i.i.d. exponential random variables with parameter θ conditional on θ. They model θ as having the gamma distribution with parameters 1 and 2. Now, suppose that they are about to observe $n = 3$ lifetimes, and they will use the posterior mean of θ as an estimator. According to Theorem 7.3.4, the posterior distribution of θ will be the gamma distribution with parameters $1 + 3 = 4$ and $2 + \sum_{i=1}^{3} X_i$. The posterior mean will then be $\hat{\theta} = 4/(2 + \sum_{i=1}^{3} X_i)$.

Prior to observing the three lifetimes, the company may want to know how likely it is that $\hat{\theta}$ will be close to θ. For example, they may want to compute $\Pr(|\hat{\theta} - \theta| < 0.1)$. In addition, other interested parties such as customers might be interested in how close the estimator is going to be to θ. But these others might not wish to assign the same prior distribution to θ. Indeed, some of them may wish to assign no prior distribution at all. We shall soon see that all of these people will find it useful to determine the sampling distribution of $\hat{\theta}$. What they do with that sampling distribution will differ, but they will all be able to make use of the sampling distribution. ◀

In Example 8.1.3, after the company observes the three lifetimes, they will be interested only in the posterior distribution of θ. They could then compute the posterior probability that $|\hat{\theta} - \theta| < 0.1$. However, before the sample is taken, both $\hat{\theta}$ and θ are random and $\Pr(|\hat{\theta} - \theta| < 0.1)$ involves the joint distribution of $\hat{\theta}$ and θ. The sampling distribution is merely the conditional distribution of $\hat{\theta}$ given θ. Hence, the law of total probability says that

$$\Pr(|\hat{\theta} - \theta| < 0.1) = E\left[\Pr(|\hat{\theta} - \theta| < 0.1|\theta)\right].$$

In this way, the company makes use of the sampling distribution of $\hat{\theta}$ as an intermediate calculation on the way to computing $\Pr(|\hat{\theta} - \theta| < 0.1)$.

Example 8.1.4

Lifetimes of Electronic Components. In Example 8.1.3, the sampling distribution of $\hat{\theta}$ does not have a name, but it is easy to see that $\hat{\theta}$ is a monotone function of the statistic $T = \sum_{i=1}^{3} X_i$ that has the gamma distribution with parameters 3 and θ (conditional on θ). So, we can compute the c.d.f. $F(\cdot|\theta)$ for the sampling distribution of $\hat{\theta}$ from the c.d.f. $G(\cdot|\theta)$ of the distribution of T. Argue as follows. For $t > 0$,

$$\begin{aligned} F(t|\theta) &= \Pr(\hat{\theta} \le t|\theta) \\ &= \Pr\left(\frac{4}{2+T} \le t \,\middle|\, \theta\right) \\ &= \Pr\left(T \ge \frac{4}{t} - 2 \,\middle|\, \theta\right) \\ &= 1 - G\left(\frac{4}{t} - 2 \,\middle|\, \theta\right). \end{aligned}$$

For $t \le 0$, $F(t|\theta) = 0$. Most statistical computer packages include the function G, which is the c.d.f. of a gamma distribution. The company can now compute, for each θ,

$$\Pr(|\hat{\theta} - \theta| < 0.1|\theta) = F(\theta + 0.1|\theta) - F(\theta - 0.1|\theta). \tag{8.1.1}$$

Figure 8.1 shows a graph of this probability as a function of θ. To complete the calculation of $\Pr(|\hat{\theta} - \theta| < 0.1)$, we must integrate (8.1.1) with respect to the distribution of θ, that is, the gamma distribution with parameters 1 and 2. This integral cannot be performed in closed form and requires a numerical approximation. One such approximation would be a simulation, which will be discussed in Chapter 12. In this example, the approximation yields $\Pr(|\hat{\theta} - \theta| < 0.1) \approx 0.478$.

Also included in Fig. 8.1 is the calculation of $\Pr(|\hat{\theta} - \theta| < 0.1|\theta)$ using $\hat{\theta} = 3/T$, the M.L.E. of θ. The sampling distribution of the M.L.E. can be derived in Exercise 9 at the end of this section. Notice that the posterior mean has higher probability of being close to θ than does the M.L.E. when θ is near the mean of the prior distribution. When θ is far from the prior mean, the M.L.E. has higher probability of being close to θ. ◀

Another case in which the sampling distribution of an estimator is needed is when the statistician must decide which one of two or more available experiments should be performed in order to obtain the best estimator of θ. For example, if she must choose which sample size to use for an experiment, then she will typically base her decision on the sampling distributions of the different estimators that might be used for each sample size.

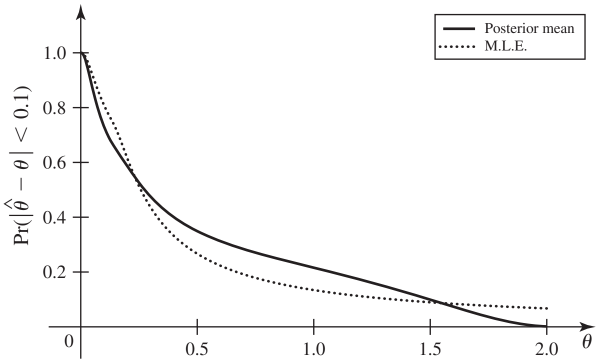


Figure 8.1 Plot of $\Pr(|\hat{\theta} - \theta| < 0.1|\theta)$ for both $\hat{\theta}$ equal to the posterior mean and $\hat{\theta}$ equal to the M.L.E. in Example 8.1.4.

As mentioned at the end of Example 8.1.3, there are statisticians who do not wish to assign a prior distribution to θ. Those statisticians would not be able to calculate a posterior distribution for θ. Instead, they would base all of their statistical inferences on the sampling distribution of whatever estimators they chose. For example, a statistician who chose to use the M.L.E. of θ in Example 8.1.4 would need to deal with the entire curve in Fig. 8.1 corresponding to the M.L.E. in order to decide how likely it is that the M.L.E. will be closer to θ than 0.1. Alternatively, she might choose a different measure of how close the M.L.E. is to θ.

Example 8.1.5 Lifetimes of Electronic Components. Suppose that a statistician chooses to estimate θ by the M.L.E., $\hat{\theta} = 3/T$ instead of the posterior mean in Example 8.1.4. This statistician may not find the graph in Fig. 8.1 very useful unless she can decide which θ values are most important to consider. Instead of calculating $\Pr(|\hat{\theta} - \theta| < 0.1|\theta)$, she might compute

$$\Pr\left(\left|\frac{\hat{\theta}}{\theta} - 1\right| < 0.1\middle|\,\theta\right). \tag{8.1.2}$$

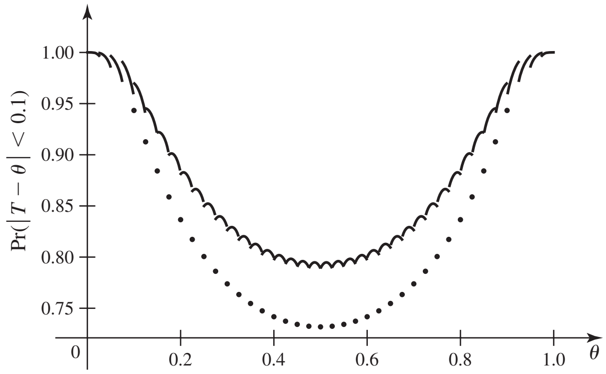
This is the probability that $\hat{\theta}$ is within 10% of the value of θ. The probability in (8.1.2) could be computed from the sampling distribution of the M.L.E. Alternatively, one can notice that $\hat{\theta}/\theta = 3/(\theta T)$, and the distribution of θT is the gamma distribution with parameters 3 and 1. Hence, $\hat{\theta}/\theta$ has a distribution that does not depend on θ. It follows that $\Pr(|\hat{\theta}/\theta - 1| < 0.1|\theta)$ is the same number for all θ. In the notation of Example 8.1.4, the c.d.f. of θT is $G(\cdot|1)$, and hence

$$\begin{aligned}
\Pr\left(\left|\frac{\hat{\theta}}{\theta} - 1\right| < 0.1\middle|\,\theta\right) &= \Pr\left(\left|\frac{3}{\theta T} - 1\right| < 0.1\middle|\,\theta\right)\\
&= \Pr\left(0.9 < \frac{3}{\theta T} < 1.1\middle|\,\theta\right)\\
&= \Pr(2.73 < \theta T < 3.33|\theta)\\
&= G(3.33|1) - G(2.73|1) = 0.134.
\end{aligned}$$

The statistician can now claim that the probability is 0.134 that the M.L.E. of θ will be within 10% of the value of θ, no matter what θ is. ◀

The random variable $\hat{\theta}/\theta$ in Example 8.1.5 is an example of a *pivotal quantity,* which will be defined and used extensively in Sec. 8.5.

Figure 8.2 Plot of $\Pr(|T - \theta| < 0.1|\theta)$ in Example 8.1.6.



Example 8.1.6

A Clinical Trial. In Example 8.1.1, we found the sampling distribution of T, the proportion of patients without relapse in the imipramine group. Using that distribution, we can draw a plot similar to that in Fig. 8.1. That is, for each θ, we can compute $\Pr(|T - \theta| < 0.1|\theta)$. The plot appears in Fig. 8.2. The jumps and cyclic nature of the plot are due the discreteness of the distribution of T. The smallest probability is 0.7318 at $\theta = 0.5$. (The isolated points that appear below the main part of the graph at θ equal to each multiple of 1/40 would appear equally far above the main part of the graph, if we had plotted $\Pr(|T - \theta| \le 0.1|\theta)$ instead of $\Pr(|T - \theta| < 0.1|\theta)$.) ◄

Summary

The sampling distribution of an estimator $\hat{\theta}$ is the conditional distribution of the estimator given the parameter. The sampling distribution can be used as an intermediate calculation in assessing the properties of a Bayes estimator prior to observing data. More commonly, the sampling distribution is used by those statisticians who prefer not to use prior and posterior distributions. For example, before the sample has been taken, the statistician can use the sampling distribution of $\hat{\theta}$ to calculate the probability that $\hat{\theta}$ will be close to θ. If this probability is high for every possible value of θ, then the statistician can feel confident that the observed value of $\hat{\theta}$ will be close to θ. After the data are observed and a particular estimate is obtained, the statistician would like to continue feeling confident that the particular estimate is likely to be close to θ, even though explicit posterior probabilities cannot be given. It is not always safe to draw such a conclusion, however, as we shall illustrate at the end of Example 8.5.11.

Exercises

1. Suppose that a random sample $X_1, \ldots, X_n$ is to be taken from the uniform distribution on the interval $[0, \theta]$ and that θ is unknown. How large a random sample must be taken in order that

$$\Pr\big(|\max\{X_1, \ldots, X_n\} - \theta| \le 0.1\theta\big) \ge 0.95,$$

for all possible θ?

2. Suppose that a random sample is to be taken from the normal distribution with unknown mean θ and standard deviation 2. How large a random sample must be taken in order that $E_\theta(|\overline{X}_n - \theta|^2) \le 0.1$ for every possible value of θ?

3. For the conditions of Exercise 2, how large a random sample must be taken in order that $E_\theta(|\overline{X}_n - \theta|) \le 0.1$ for every possible value of θ?

4. For the conditions of Exercise 2, how large a random sample must be taken in order that $\Pr(|\overline{X}_n - \theta| \le 0.1) \ge 0.95$ for every possible value of θ?

5. Suppose that a random sample is to be taken from the Bernoulli distribution with unknown parameter p. Suppose also that it is believed that the value of p is in the neighborhood of 0.2. How large a random sample must be taken in order that $\Pr(|\overline{X}_n - p| \le 0.1) \ge 0.75$ when $p = 0.2$?

6. For the conditions of Exercise 5, use the central limit theorem in Sec. 6.3 to find approximately the size of a random sample that must be taken in order that $\Pr(|\overline{X}_n - p| \le 0.1) \ge 0.95$ when $p = 0.2$.

7. For the conditions of Exercise 5, how large a random sample must be taken in order that $E_p(|\overline{X}_n - p|^2) \le 0.01$ when $p = 0.2$?

8. For the conditions of Exercise 5, how large a random sample must be taken in order that $E_p(|\overline{X}_n - p|^2) \le 0.01$ for every possible value of p $(0 \le p \le 1)$?

9. Let $X_1, \ldots, X_n$ be a random sample from the exponential distribution with parameter θ. Find the c.d.f. for the sampling distribution of the M.L.E. of θ. (The M.L.E. itself was found in Exercise 7 in Sec. 7.5.)

8.2 The Chi-Square Distributions

The family of chi-square (χ^2) distributions is a subcollection of the family of gamma distributions. These special gamma distributions arise as sampling distributions of variance estimators based on random samples from a normal distribution.

Definition of the Distributions

Example 8.2.1 M.L.E. of the Variance of a Normal Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with known mean μ and unknown variance σ^2. The M.L.E. of σ^2 is found in Exercise 6 in Sec. 7.5. It is

$$\widehat{\sigma_0^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

The distributions of $\widehat{\sigma_0^2}$ and $\widehat{\sigma_0^2}/\sigma^2$ are useful in several statistical problems, and we shall derive them in this section. ◀

In this section, we shall introduce and discuss a particular class of gamma distributions known as the chi-square (χ^2) distributions. These distributions, which are closely related to random samples from a normal distribution, are widely applied in the field of statistics. In the remainder of this book, we shall see how they are applied in many important problems of statistical inference. In this section, we shall present the definition of the χ^2 distributions and some of their basic mathematical properties.

Definition 8.2.1 χ^2 Distributions. For each positive number m, the gamma distribution with parameters $\alpha = m/2$ and $\beta = 1/2$ is called the χ^2 *distribution with m degrees of freedom*. (See Definition 5.7.2 for the definition of the gamma distribution with parameters α and β.)

It is common to restrict the degrees of freedom m in Definition 8.2.1 to be an integer. However, there are situations in which it will be useful for the degrees of freedom to not be integers, so we will not make that restriction.

If a random variable X has the χ^2 distribution with m degrees of freedom, it follows from Eq. (5.7.13) that the p.d.f. of X for $x > 0$ is

$$f(x) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{(m/2)-1} e^{-x/2}. \quad (8.2.1)$$

Also, $f(x) = 0$ for $x \le 0$.

A short table of p quantiles for the χ^2 distribution for various values of p and various degrees of freedom is given at the end of this book. Most statistical software packages include functions to compute the c.d.f. and the quantile function of an arbitrary χ^2 distribution.

It follows from Definition 8.2.1, and it can be seen from Eq. (8.2.1), that the χ^2 distribution with two degrees of freedom is the exponential distribution with parameter 1/2 or, equivalently, the exponential distribution for which the mean is 2. Thus, the following three distributions are all the same: the gamma distribution with parameters $\alpha = 1$ and $\beta = 1/2$, the χ^2 distribution with two degrees of freedom, and the exponential distribution for which the mean is 2.

Properties of the Distributions

The means and variances of χ^2 distributions follow immediately from Theorem 5.7.5, and are given here without proof.

Theorem 8.2.1 Mean and Variance. If a random variable X has the χ^2 distribution with m degrees of freedom, then $E(X) = m$ and $\text{Var}(X) = 2m$. ■

Furthermore, it follows from the moment generating function given in Eq. (5.7.15) that the m.g.f. of X is

$$\psi(t) = \left(\frac{1}{1-2t}\right)^{m/2} \quad \text{for } t < \frac{1}{2}.$$

The additivity property of the χ^2 distribution, which is presented without proof in the next theorem, follows directly from Theorem 5.7.7.

Theorem 8.2.2 If the random variables $X_1, \ldots, X_k$ are independent and if X_i has the χ^2 distribution with m_i degrees of freedom ($i = 1, \ldots, k$), then the sum $X_1 + \cdots + X_k$ has the χ^2 distribution with $m_1 + \cdots + m_k$ degrees of freedom. ■

We shall now establish the basic relation between the χ^2 distributions and the standard normal distribution.

Theorem 8.2.3 Let X have the standard normal distribution. Then the random variable $Y = X^2$ has the χ^2 distribution with one degree of freedom.

Proof Let $f(y)$ and $F(y)$ denote, respectively, the p.d.f. and the c.d.f. of Y. Also, since X has the standard normal distribution, we shall let $\phi(x)$ and $\Phi(x)$ denote the p.d.f. and the c.d.f. of X. Then for $y > 0$,

$$\begin{aligned} F(y) &= \Pr(Y \le y) = \Pr(X^2 \le y) = \Pr(-y^{1/2} \le X \le y^{1/2}) \\ &= \Phi(y^{1/2}) - \Phi(-y^{1/2}). \end{aligned}$$

Since $f(y) = F'(y)$ and $\phi(x) = \Phi'(x)$, it follows from the chain rule for derivatives that

$$f(y) = \phi(y^{1/2})\left(\frac{1}{2}y^{-1/2}\right) + \phi(-y^{1/2})\left(\frac{1}{2}y^{-1/2}\right).$$

Furthermore, since $\phi(y^{1/2}) = \phi(-y^{1/2}) = (2\pi)^{-1/2}e^{-y/2}$, it now follows that

$$f(y) = \frac{1}{(2\pi)^{1/2}} y^{-1/2} e^{-y/2} \quad \text{for } y > 0.$$

By comparing this equation with Eq. (8.2.1), it is seen that the p.d.f. of Y is indeed the p.d.f. of the χ^2 distribution with one degree of freedom. ■

We can now combine Theorem 8.2.3 with Theorem 8.2.2 to obtain the following result, which provides the main reason that the χ^2 distribution is important in statistics.

Corollary 8.2.1 If the random variables $X_1, \ldots, X_m$ are i.i.d. with the standard normal distribution, then the sum of squares $X_1^2 + \cdots + X_m^2$ has the χ^2 distribution with m degrees of freedom. ■

Example 8.2.2 M.L.E. of the Variance of a Normal Distribution. In Example 8.2.1, the random variables $Z_i = (X_i - \mu)/\sigma$ for $i = 1, \ldots, n$ form a random sample from the standard normal distribution. It follows from Corollary 8.2.1 that the distribution of $\sum_{i=1}^n Z_i^2$ is the χ^2 distribution with n degrees of freedom. It is easy to see that $\sum_{i=1}^n Z_i^2$ is precisely the same as $n\widehat{\sigma_0^2}/\sigma^2$, which appears in Example 8.2.1. So the distribution of $n\widehat{\sigma_0^2}/\sigma^2$ is the χ^2 distribution with n degrees of freedom. The reader should also be able to see that the distribution of $\widehat{\sigma_0^2}$ itself is the gamma distribution with parameters $n/2$ and $n/(2\sigma^2)$ (Exercise 13). ◀

Example 8.2.3 Acid Concentration in Cheese. Moore and McCabe (1999, p. D-1) describe an experiment conducted in Australia to study the relationship between taste and the chemical composition of cheese. One chemical whose concentration can affect taste is lactic acid. Cheese manufacturers who want to establish a loyal customer base would like the taste to be about the same each time a customer purchases the cheese. The variation in concentrations of chemicals like lactic acid can lead to variation in the taste of cheese. Suppose that we model the concentration of lactic acid in several chunks of cheese as independent normal random variables with mean μ and variance σ^2. We are interested in how much these concentrations differ from the value μ. Let $X_1, \ldots, X_k$ be the concentrations in k chunks, and let $Z_i = (X_i - \mu)/\sigma$. Then

$$Y = \frac{1}{k}\sum_{i=1}^{k} |X_i - \mu|^2 = \frac{\sigma^2}{k}\sum_{i=1}^{k} Z_i^2$$

is one measure of how much the k concentrations differ from μ. Suppose that a difference of u or more in lactic acid concentration is enough to cause a noticeable difference in taste. We might then wish to calculate $\Pr(Y \le u^2)$. According to Corollary 8.2.1, the distribution of $W = kY/\sigma^2$ is χ^2 with k degrees of freedom. Hence, $\Pr(Y \le u^2) = \Pr(W \le ku^2/\sigma^2)$.

For example, suppose that $\sigma^2 = 0.09$, and we are interested in $k = 10$ cheese chunks. Furthermore, suppose that $u = 0.3$ is the critical difference of interest. We

can write

$$\Pr(Y \le 0.3^2) = \Pr\left(W \le \frac{10 \times 0.09}{0.09}\right) = \Pr(W \le 10). \tag{8.2.2}$$

Using the table of quantiles of the χ^2 distribution with 10 degrees of freedom, we see that 10 is between the 0.5 and 0.6 quantiles. In fact, the probability in Eq. (8.2.2) can be found by computer software to equal 0.56, so there is a 44 percent chance that the average squared difference between lactic acid concentration and mean concentration in 10 chunks will be more than the desired amount. If this probability is too large, the manufacturer might wish to invest some effort in reducing the variance of lactic acid concentration. ◄

Summary

The chi-square distribution with n degrees of freedom is the same as the gamma distribution with parameters $m/2$ and $1/2$. It is the distribution of the sum of squares of a sample of m independent standard normal random variables. The mean of the χ^2 distribution with m degrees of freedom is m, and the variance is $2m$.

Exercises

1. Suppose that we will sample 20 chunks of cheese in Example 8.2.3. Let $T = \sum_{i=1}^{20}(X_i - \mu)^2/20$, where X_i is the concentration of lactic acid in the ith chunk. Assume that $\sigma^2 = 0.09$. What number c satisfies $\Pr(T \le c) = 0.9$?

2. Find the mode of the χ^2 distribution with m degrees of freedom ($m = 1, 2, \ldots$).

3. Sketch the p.d.f. of the χ^2 distribution with m degrees of freedom for each of the following values of m. Locate the mean, the median, and the mode on each sketch. **(a)** $m = 1$; **(b)** $m = 2$; **(c)** $m = 3$; **(d)** $m = 4$.

4. Suppose that a point (X, Y) is to be chosen at random in the xy-plane, where X and Y are independent random variables and each has the standard normal distribution. If a circle is drawn in the xy-plane with its center at the origin, what is the radius of the smallest circle that can be chosen in order for there to be probability 0.99 that the point (X, Y) will lie inside the circle?

5. Suppose that a point (X, Y, Z) is to be chosen at random in three-dimensional space, where X, Y, and Z are independent random variables and each has the standard normal distribution. What is the probability that the distance from the origin to the point will be less than 1 unit?

6. When the motion of a microscopic particle in a liquid or a gas is observed, it is seen that the motion is irregular because the particle collides frequently with other particles. The probability model for this motion, which is called *Brownian motion,* is as follows: A coordinate system is chosen in the liquid or gas. Suppose that the particle is at the origin of this coordinate system at time $t = 0$, and let (X, Y, Z) denote the coordinates of the particle at any time $t > 0$. The random variables X, Y, and Z are i.i.d., and each of them has the normal distribution with mean 0 and variance $\sigma^2 t$. Find the probability that at time $t = 2$ the particle will lie within a sphere whose center is at the origin and whose radius is 4σ.

7. Suppose that the random variables $X_1, \ldots, X_n$ are independent, and each random variable X_i has a continuous c.d.f. F_i. Also, let the random variable Y be defined by the relation $Y = -2\sum_{i=1}^{n} \log F_i(X_i)$. Show that Y has the χ^2 distribution with $2n$ degrees of freedom.

8. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval [0, 1], and let W denote the range of the sample, as defined in Example 3.9.7. Also, let $g_n(x)$ denote the p.d.f. of the random variable $2n(1 - W)$, and let $g(x)$ denote the p.d.f. of the χ^2 distribution with four degrees of freedom. Show that

$$\lim_{n\to\infty} g_n(x) = g(x) \quad \text{for } x > 0.$$

9. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Find the distribution of

$$\frac{n(\overline{X}_n - \mu)^2}{\sigma^2}.$$

10. Suppose that six random variables $X_1, \ldots, X_6$ form a random sample from the standard normal distribution, and let

$$Y = (X_1 + X_2 + X_3)^2 + (X_4 + X_5 + X_6)^2.$$

Determine a value of c such that the random variable cY will have a χ^2 distribution.

11. If a random variable X has the χ^2 distribution with m degrees of freedom, then the distribution of $X^{1/2}$ is called a *chi* (χ) *distribution with m degrees of freedom*. Determine the mean of this distribution.

12. Consider again the situation described in Example 8.2.3. How small would σ^2 need to be in order for $\Pr(Y \le 0.09) \ge 0.9$?

13. Prove that the distribution of $\widehat{\sigma_0^2}$ in Examples 8.2.1 and 8.2.2 is the gamma distribution with parameters $n/2$ and $n/(2\sigma^2)$.

8.3 Joint Distribution of the Sample Mean and Sample Variance

Suppose that our data form a random sample from a normal distribution. The sample mean $\hat{\mu}$ and sample variance $\widehat{\sigma^2}$ are important statistics that are computed in order to estimate the parameters of the normal distribution. Their marginal distributions help us to understand how good each of them is as an estimator of the corresponding parameter. However, the marginal distribution of $\hat{\mu}$ depends on σ. The joint distribution of $\hat{\mu}$ and $\widehat{\sigma^2}$ will allow us to make inferences about μ without reference to σ.

Independence of the Sample Mean and Sample Variance

Example 8.3.1

Rain from Seeded Clouds. Simpson, Olsen, and Eden (1975) describe an experiment in which a random sample of 26 clouds were seeded with silver nitrate to see if they produced more rain than unseeded clouds. Suppose that, on a log scale, unseeded clouds typically produced a mean rainfall of 4. In comparing the mean of the seeded clouds to the unseeded mean, one might naturally see how far the average log-rainfall of the seeded clouds $\hat{\mu}$ is from 4. But the variation in rainfall within the sample is also important. For example, if one compared two different samples of seeded clouds, one would expect the average rainfalls in the two samples to be different just due to variation between clouds. In order to be confident that seeding the clouds really produced more rain, we would want the average log-rainfall to exceed 4 by a large amount compared to the variation between samples, which is closely related to the variation within samples. Since we do not know the variance for seeded clouds, we compute the sample variance $\widehat{\sigma^2}$. Comparing $\hat{\mu} - 4$ to $\widehat{\sigma^2}$ requires us to consider the joint distribution of the sample mean and the sample variance. ◄

Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Then, as was shown in Example 7.5.6, the M.L.E.'s of μ and σ^2 are the sample mean $\overline{X}_n$ and the sample variance $(1/n)\sum_{i=1}^n (X_i - \overline{X}_n)^2$. In this section, we shall derive the joint distribution of these two estimators.

We already know from Corollary 5.6.2 that the sample mean itself has the normal distribution with mean μ and variance σ^2/n. We shall establish the noteworthy property that the sample mean and the sample variance are independent random variables, even though both are functions of the same random variables $X_1, \dots, X_n$. Furthermore, we shall show that, except for a scale factor, the sample variance has the χ^2 distribution with $n-1$ degrees of freedom. More precisely, we shall show that the random variable $\sum_{i=1}^n (X_i - \overline{X}_n)^2/\sigma^2$ has the χ^2 distribution with $n-1$ degrees

of freedom. This result is also a rather striking property of random samples from a normal distribution, as the following discussion indicates.

Because the random variables $X_1, \dots, X_n$ are independent, and because each has the normal distribution with mean μ and variance σ^2, the random variables $(X_1 - \mu)/\sigma, \dots, (X_n - \mu)/\sigma$ are also independent, and each of these variables has the standard normal distribution. It follows from Corollary 8.2.1 that the sum of their squares $\sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2$ has the χ^2 distribution with n degrees of freedom. Hence, the striking property mentioned in the previous paragraph is that if the population mean μ is replaced by the sample mean $\overline{X}_n$ in this sum of squares, the effect is simply to reduce the degrees of freedom in the χ^2 distribution from n to $n - 1$. In summary, we shall establish the following theorem.

Theorem 8.3.1 Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Then the sample mean $\overline{X}_n$ and the sample variance $(1/n)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ are independent random variables, $\overline{X}_n$ has the normal distribution with mean μ and variance σ^2/n, and $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2/\sigma^2$ has the χ^2 distribution with $n - 1$ degrees of freedom.

Furthermore, it can be shown that the sample mean and the sample variance are independent *only* when the random sample is drawn from a normal distribution. We shall not consider this result further in this book. However, it does emphasize the fact that the independence of the sample mean and the sample variance is indeed a noteworthy property of samples from a normal distribution.

The proof of Theorem 8.3.1 makes use of transformations of several variables as described in Sec. 3.9 and of the properties of orthogonal matrices. The proof appears at the end of this section.

Example 8.3.2 Rain from Seeded Clouds. Figure 8.3 is a histogram of the logarithms of the rainfalls from the seeded clouds in Example 8.3.1. Suppose that these logarithms $X_1, \dots, X_{26}$ are modeled as i.i.d. normal random variables with mean μ and variance σ^2. If we are interested in how much variation there is in rainfall among the seeded clouds, we can compute the sample variance $\widehat{\sigma^2} = \sum_{i=1}^{26}(X_i - \overline{X}_n)^2/26$. The distribution of $U = 26\widehat{\sigma^2}/\sigma^2$ is the χ^2 distribution with 25 degrees of freedom. We can use this distribution to tell us how likely it is that $\widehat{\sigma^2}$ will overestimate or underestimate σ^2 by various amounts. For example, the χ^2 table in this book says that the 0.25 quantile of the χ^2 distribution with 25 degrees of freedom is 19.94, so $\Pr(U \le 19.94) = 0.25$.

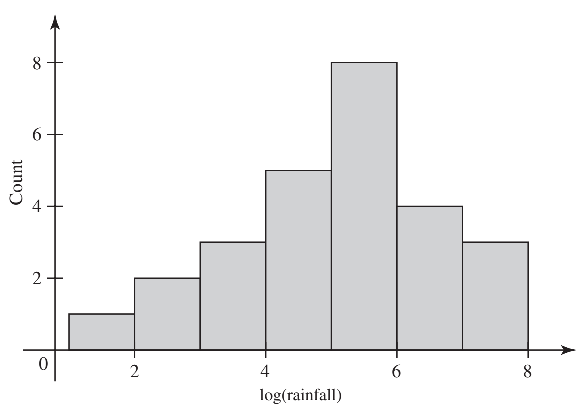


Figure 8.3 Histogram of log-rainfalls from seeded clouds.

It follows that

$$0.25 = \Pr\left(\frac{\widehat{\sigma^2}}{\sigma^2} \le \frac{19.94}{26}\right) = \Pr(\widehat{\sigma^2} \le 0.77\sigma^2). \tag{8.3.1}$$

That is, there is probability 0.25 that $\widehat{\sigma^2}$ will underestimate σ^2 by 23 percent or more. The observed value of $\widehat{\sigma^2}$ is 2.460 in this example. The probability calculated in Eq. (8.3.1) has nothing to do with how far 2.460 is from σ^2. Eq. (8.3.1) tells us the probability (prior to observing the data) that $\widehat{\sigma^2}$ would be at least 23% below σ^2. ◀

Estimation of the Mean and Standard Deviation

We shall assume that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown standard deviation σ. Also, as usual, we shall denote the M.L.E.'s of μ and σ by $\hat{\mu}$ and $\hat{\sigma}$. Thus,

$$\hat{\mu} = \overline{X}_n \quad \text{and} \quad \hat{\sigma} = \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right)^{1/2}.$$

Notice that $\hat{\sigma}^2 = \widehat{\sigma^2}$, the M.L.E. of σ^2. For the remainder of this book, when referring to the M.L.E. of σ^2, we shall use whichever symbol $\hat{\sigma}^2$ or $\widehat{\sigma^2}$ is more convenient. As an illustration of the application of Theorem 8.3.1, we shall now determine the smallest possible sample size n such that the following relation will be satisfied:

$$\Pr\left(|\hat{\mu} - \mu| \le \frac{1}{5}\sigma \quad \text{and} \quad |\hat{\sigma} - \sigma| \le \frac{1}{5}\sigma\right) \ge \frac{1}{2}. \tag{8.3.2}$$

In other words, we shall determine the minimum sample size n for which the probability will be at least 1/2 that neither $\hat{\mu}$ nor $\hat{\sigma}$ will differ from the unknown value it is estimating by more than $(1/5)\sigma$.

Because of the independence of $\hat{\mu}$ and $\hat{\sigma}$, the relation (8.3.2) can be rewritten as follows:

$$\Pr\left(|\hat{\mu} - \mu| < \frac{1}{5}\sigma\right)\Pr\left(|\hat{\sigma} - \sigma| < \frac{1}{5}\sigma\right) \ge \frac{1}{2}. \tag{8.3.3}$$

If we let p_1 denote the first probability on the left side of the relation (8.3.3), and let U be a random variable that has the standard normal distribution, this probability can be written in the following form:

$$p_1 = \Pr\left(\frac{\sqrt{n}|\hat{\mu} - \mu|}{\sigma} < \frac{1}{5}\sqrt{n}\right) = \Pr\left(|U| < \frac{1}{5}\sqrt{n}\right).$$

Similarly, if we let p_2 denote the second probability on the left side of the relation (8.3.3), and let $V = n\hat{\sigma}^2/\sigma^2$, this probability can be written in the following form:

$$\begin{aligned} p_2 &= \Pr\left(0.8 < \frac{\hat{\sigma}}{\sigma} < 1.2\right) = \Pr\left(0.64n < \frac{n\hat{\sigma}^2}{\sigma^2} < 1.44n\right) \\ &= \Pr(0.64n < V < 1.44n). \end{aligned}$$

By Theorem 8.3.1, the random variable V has the χ^2 distribution with $n - 1$ degrees of freedom.

For each specific value of n, the values of p_1 and p_2 can be found, at least approximately, from the table of the standard normal distribution and the table of the χ^2 distribution given at the end of this book. In particular, after various values

of n have been tried, it will be found that for $n = 21$ the values of p_1 and p_2 are $p_1 = 0.64$ and $p_2 = 0.78$. Hence, $p_1 p_2 = 0.50$, and it follows that the relation (8.3.2) will be satisfied for $n = 21$.

Proof of Theorem 8.3.1

We already knew from Corollary 5.6.2 that the distribution of the sample mean was as stated in Theorem 8.3.1. What remains to prove is the stated distribution of the sample variance and the independence of the sample mean and sample variance.

Orthogonal Matrices

We begin with some properties of orthogonal matrices that are essential for the proof.

Definition 8.3.1 Orthogonal Matrix. It is said that an $n \times n$ matrix $\boldsymbol{A}$ is *orthogonal* if $\boldsymbol{A}^{-1} = \boldsymbol{A}'$, where $\boldsymbol{A}'$ is the transpose of $\boldsymbol{A}$.

In other words, a matrix $\boldsymbol{A}$ is orthogonal if and only if $\mathbf{AA}' = \mathbf{A}'\mathbf{A} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix. It follows from this latter property that a matrix is orthogonal if and only if the sum of the squares of the elements in each row is 1 and the sum of the products of the corresponding elements in every pair of different rows is 0. Alternatively, a matrix is orthogonal if and only if the sum of the squares of the elements in each column is 1 and the sum of the products of the corresponding elements in every pair of different columns is 0.

Properties of Orthogonal Matrices We shall now derive two important properties of orthogonal matrices.

Theorem 8.3.2 Determinant is 1. If $\boldsymbol{A}$ is orthogonal, then $|\det \boldsymbol{A}| = 1$.

Proof To prove this result, it should be recalled that $\det \boldsymbol{A} = \det \boldsymbol{A}'$ for every square matrix $\boldsymbol{A}$. Also recall that $\det \boldsymbol{AB} = (\det \boldsymbol{A})(\det \boldsymbol{B})$ for square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Therefore,

$$\det(\boldsymbol{AA}') = (\det \boldsymbol{A})(\det \boldsymbol{A}') = (\det \boldsymbol{A})^2.$$

Also, if $\boldsymbol{A}$ is orthogonal, then $\boldsymbol{AA}' = \boldsymbol{I}$, and it follows that

$$\det(\boldsymbol{AA}') = \det \boldsymbol{I} = 1.$$

Hence $(\det \boldsymbol{A})^2 = 1$ or, equivalently, $|\det \boldsymbol{A}| = 1$. ■

Theorem 8.3.3 Squared Length Is Preserved. Consider two n-dimensional random vectors

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \tag{8.3.4}$$

and suppose that $\boldsymbol{Y} = \boldsymbol{AX}$, where $\boldsymbol{A}$ is an orthogonal matrix. Then

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2. \tag{8.3.5}$$

Proof This result follows from the fact that $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}$, because

$$\sum_{i=1}^{N} Y_i^2 = \boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{X} = \sum_{i=1}^{n} X_i^2. \qquad \blacksquare$$

Multiplication of a vector $\boldsymbol{X}$ by an orthogonal matrix $\boldsymbol{A}$ corresponds to a rotation of $\boldsymbol{X}$ in n-dimensional space possibly followed by changing the signs of some coordinates. Neither of these operations can change the length of the original vector $\boldsymbol{X}$, and that length equals $(\sum_{i=1}^{n} X_i^2)^{1/2}$.

Together, these two properties of orthogonal matrices imply that if a random vector $\boldsymbol{Y}$ is obtained from a random vector $\boldsymbol{X}$ by an *orthogonal* linear transformation $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, then the absolute value of the Jacobian of the transformation is 1 and $\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2$.

We combine Theorems 8.3.2 and 8.3.3 to obtain a useful fact about orthogonal transformations of a random sample of standard normal random variables.

Theorem 8.3.4 Suppose that the random variables, $X_1, \ldots, X_n$ are i.i.d. and each has the standard normal distribution. Suppose also that $\boldsymbol{A}$ is an orthogonal $n \times n$ matrix, and $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$. Then the random variables $Y_1, \ldots, Y_n$ are also i.i.d., each also has the standard normal distribution, and $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i^2$.

Proof The joint p.d.f. of $X_1, \ldots, X_n$ is as follows, for $-\infty < x_i < \infty$ $(i = 1, \ldots, n)$:

$$f_n(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right). \tag{8.3.6}$$

If $\boldsymbol{A}$ is an orthogonal $n \times n$ matrix, and the random variables $Y_1, \ldots, Y_n$ are defined by the relation $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, where the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are as specified in Eq. (8.3.4). This is a linear transformation, so the joint p.d.f. of $Y_1, \ldots, Y_n$ is obtained from Eq. (3.9.20) and equals

$$g_n(\boldsymbol{y}) = \frac{1}{|\det \boldsymbol{A}|} f_n(\boldsymbol{A}^{-1}\boldsymbol{y}).$$

Let $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}$. Since $\boldsymbol{A}$ is orthogonal, $|\det \boldsymbol{A}| = 1$ and $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} x_i^2$, as we just proved. So,

$$g_n(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} y_i^2\right). \tag{8.3.7}$$

It can be seen from Eq. (8.3.7) that the joint p.d.f. of $Y_1, \ldots, Y_n$ is exactly the same as the joint p.d.f. of $X_1, \ldots, X_n$. $\blacksquare$

Proof of Theorem 8.3.1

Random Samples from the Standard Normal Distribution We shall begin by proving Theorem 8.3.1 under the assumption that $X_1, \ldots, X_n$ form a random sample from the standard normal distribution. Consider the n-dimensional row vector $\boldsymbol{u}$, in which each of the n components has the value $1/\sqrt{n}$:

$$\boldsymbol{u} = \left[\frac{1}{\sqrt{n}} \cdots \frac{1}{\sqrt{n}}\right]. \tag{8.3.8}$$

Since the sum of the squares of the n components of the vector $\boldsymbol{u}$ is 1, it is possible to construct an orthogonal matrix $\boldsymbol{A}$ such that the components of the vector $\boldsymbol{u}$ form

the first row of $\boldsymbol{A}$. This construction, called the *Gram-Schmidt method*, is described in textbooks on linear algebra such as Cullen (1972) and will not be discussed here. We shall assume that such a matrix $\boldsymbol{A}$ has been constructed, and we shall again define the random variables $Y_1, \dots, Y_n$ by the transformation $\boldsymbol{Y} = \boldsymbol{AX}$.

Since the components of $\boldsymbol{u}$ form the first row of $\boldsymbol{A}$, it follows that

$$Y_1 = \boldsymbol{uX} = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} X_i = \sqrt{n}\, \overline{X}_n. \tag{8.3.9}$$

Furthermore, by Theorem 8.3.4, $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i^2$. Therefore,

$$\sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2 = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

We have thus obtained the relation

$$\sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2. \tag{8.3.10}$$

It is known from Theorem 8.3.4 that the random variables $Y_1, \dots, Y_n$ are independent. Therefore, the two random variables Y_1 and $\sum_{i=2}^{n} Y_i^2$ are independent, and it follows from Eqs. (8.3.9) and (8.3.10) that $\overline{X}_n$ and $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ are independent. Furthermore, it is known from Theorem 8.3.4 that the $n-1$ random variables $Y_2, \dots, Y_n$ are i.i.d., and that each of these random variables has the standard normal distribution. Hence, by Corollary 8.2.1 the random variable $\sum_{i=2}^{n} Y_i^2$ has the χ^2 distribution with $n-1$ degrees of freedom. It follows from Eq. (8.3.10) that $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ also has the χ^2 distribution with $n-1$ degrees of freedom.

Random Samples from an Arbitrary Normal Distribution Thus far, in proving Theorem 8.3.1, we have considered only random samples from the standard normal distribution. Suppose now that the random variables $X_1, \dots, X_n$ form a random sample from an arbitrary normal distribution with mean μ and variance σ^2.

If we let $Z_i = (X_i - \mu)/\sigma$ for $i = 1, \dots, n$, then the random variables $Z_1, \dots, Z_n$ are independent, and each has the standard normal distribution. In other words, the joint distribution of $Z_1, \dots, Z_n$ is the same as the joint distribution of a random sample from the standard normal distribution. It follows from the results that have just been obtained that $\overline{Z}_n$ and $\sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2$ are independent, and $\sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2$ has the χ^2 distribution with $n-1$ degrees of freedom. However, $\overline{Z}_n = (\overline{X}_n - \mu)/\sigma$ and

$$\sum_{i=1}^{n} (Z_i - \overline{Z}_n)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2. \tag{8.3.11}$$

We now conclude that the sample mean $\overline{X}_n$ and the sample variance $(1/n)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ are independent, and that the random variable on the right side of Eq. (8.3.11) has the χ^2 distribution with $n-1$ degrees of freedom. All the results stated in Theorem 8.3.1 have now been established.

Summary

Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2. Then the sample mean $\hat{\mu} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and sample variance $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ are independent random variables. Furthermore, $\hat{\mu}$ has the normal distribution with mean μ and variance σ^2/n, and $n\widehat{\sigma^2}/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom.

Exercises

1. Assume that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Show that $\widehat{\sigma^2}$ has the gamma distribution with parameters $(n-1)/2$ and $n/(2\sigma^2)$.

2. Determine whether or not each of the five following matrices is orthogonal:

a. $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ **b.** $\begin{bmatrix} 0.8 & 0 & 0.6 \\ -0.6 & 0 & 0.8 \\ 0 & -1 & 0 \end{bmatrix}$

c. $\begin{bmatrix} 0.8 & 0 & 0.6 \\ -0.6 & 0 & 0.8 \\ 0 & 0.5 & 0 \end{bmatrix}$ **d.** $\begin{bmatrix} -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \end{bmatrix}$

e. $\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}$

3.a. Construct a 2×2 orthogonal matrix for which the first row is as follows:

$$\left[\begin{matrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{matrix}\right].$$

b. Construct a 3×3 orthogonal matrix for which the first row is as follows:

$$\left[\begin{matrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{matrix}\right].$$

4. Suppose that the random variables X_1, X_2, and X_3 are i.i.d., and that each has the standard normal distribution. Also, suppose that

$$Y_1 = 0.8X_1 + 0.6X_2,$$
$$Y_2 = \sqrt{2}(0.3X_1 - 0.4X_2 - 0.5X_3),$$
$$Y_3 = \sqrt{2}(0.3X_1 - 0.4X_2 + 0.5X_3).$$

Find the joint distribution of Y_1, Y_2, and Y_3.

5. Suppose that the random variables X_1 and X_2 are independent, and that each has the normal distribution with mean μ and variance σ^2. Prove that the random variables $X_1 + X_2$ and $X_1 - X_2$ are independent.

6. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Assuming that the sample size n is 16, determine the values of the following probabilities:

a. $\Pr\left[\frac{1}{2}\sigma^2 \le \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 \le 2\sigma^2\right]$

b. $\Pr\left[\frac{1}{2}\sigma^2 \le \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 \le 2\sigma^2\right]$

7. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2, and let $\widehat{\sigma^2}$ denote the sample variance. Determine the smallest values of n for which the following relations are satisfied:

a. $\Pr\left(\frac{\widehat{\sigma^2}}{\sigma^2} \le 1.5\right) \ge 0.95$

b. $\Pr\left(|\widehat{\sigma^2} - \sigma^2| \le \frac{1}{2}\sigma^2\right) \ge 0.8$

8. Suppose that X has the χ^2 distribution with 200 degrees of freedom. Explain why the central limit theorem can be used to determine the approximate value of $\Pr(160 < X < 240)$ and find this approximate value.

9. Suppose that each of two statisticians, A and B, independently takes a random sample of 20 observations from the normal distribution with unknown mean μ and known variance 4. Suppose also that statistician A finds the sample variance in his random sample to be 3.8, and statistician B finds the sample variance in her random sample to be 9.4. For which random sample is the sample mean likely to be closer to the unknown value of μ?

8.4 The t Distributions

When our data are a sample from the normal distribution with mean μ and variance σ^2, the distribution of $Z = n^{1/2}(\hat{\mu} - \mu)/\sigma$ is the standard normal distribution, where $\hat{\mu}$ is the sample mean. If σ^2 is unknown, we can replace σ by an estimator (similar to the M.L.E.) in the formula for Z. The resulting random variable has the t distribution with $n - 1$ degrees of freedom and is useful for making inferences about μ alone even when both μ and σ^2 are unknown.

Definition of the Distributions

Example 8.4.1 Rain from Seeded Clouds. Consider the same sample of log-rainfall measurements from 26 seeded clouds from Example 8.3.2. Suppose now that we are interested in how far the sample average $\overline{X}_n$ of those measurements is from the mean μ. We know that $n^{1/2}(\overline{X}_n - \mu)/\sigma$ has the standard normal distribution, but we do not know σ. If we replace σ by an estimator $\hat{\sigma}$ such as the M.L.E., or something similar, what is the distribution of $n^{1/2}(\overline{X}_n - \mu)/\hat{\sigma}$, and how can we make use of this random variable to make inferences about μ? ◀

In this section, we shall introduce and discuss another family of distributions, called the t distributions, which are closely related to random samples from a normal distribution. The t distributions, like the χ^2 distributions, have been widely applied in important problems of statistical inference. The t distributions are also known as Student's distributions (see Student, 1908), in honor of W. S. Gosset, who published his studies of this distribution in 1908 under the pen name "Student." The distributions are defined as follows.

Definition 8.4.1 t Distributions. Consider two independent random variables Y and Z, such that Y has the χ^2 distribution with m degrees of freedom and Z has the standard normal distribution. Suppose that a random variable X is defined by the equation

$$X = \frac{Z}{\left(\frac{Y}{m}\right)^{1/2}}. \tag{8.4.1}$$

Then the distribution of X is called the t *distribution with m degrees of freedom*.

The derivation of the p.d.f. of the t distribution with m degrees of freedom makes use of the methods of Sec. 3.9 and will be given at the end of this section. But we state the result here.

Theorem 8.4.1 Probability Density Function. The p.d.f. of the t distribution with m degrees of freedom is

$$\frac{\Gamma\left(\frac{m+1}{2}\right)}{(m\pi)^{1/2}\Gamma\left(\frac{m}{2}\right)}\left(1 + \frac{x^2}{m}\right)^{-(m+1)/2} \quad \text{for } -\infty < x < \infty. \tag{8.4.2}$$

Moments of the t Distributions Although the mean of the t distribution does not exist when $m \le 1$, the mean does exist for every value of $m > 1$. Of course, whenever the mean does exist, its value is 0 because of the symmetry of the t distribution.

In general, if a random variable X has the t distribution with m degrees of freedom ($m > 1$), then it can be shown that $E(|X|^k) < \infty$ for $k < m$ and that $E(|X|^k) = \infty$ for $k \geq m$. If m is an integer, the first $m - 1$ moments of X exist, but no moments of higher order exist. It follows, therefore, that the m.g.f. of X does not exist.

It can be shown (see Exercise 1 at the end of this section) that if X has the t distribution with m degrees of freedom ($m > 2$), then $\text{Var}(X) = m/(m-2)$.

Relation to Random Samples from a Normal Distribution

Example 8.4.2 Rain from Seeded Clouds. Return to Example 8.4.1. We have already seen that $Z = n^{1/2}(\overline{X}_n - \mu)/\sigma$ has the standard normal distribution. Furthermore, Theorem 8.3.1 says that $\overline{X}_n$ (and hence Z) is independent of $Y = n\widehat{\sigma^2}/\sigma^2$, which has the χ^2 distribution with $n-1$ degrees of freedom. It follows that $Z/(Y/[n-1])^{1/2}$ has the t distribution with $n-1$ degrees of freedom. We shall show how to use this fact after stating the general version of this result. ◀

Theorem 8.4.2 Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Let $\overline{X}_n$ denote the sample mean, and define

$$\sigma' = \left[\frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n-1}\right]^{1/2}. \tag{8.4.3}$$

Then $n^{1/2}(\overline{X}_n - \mu)/\sigma'$ has the t distribution with $n-1$ degrees of freedom.

Proof Define $S_n^2 = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. Next, define $Z = n^{1/2}(\overline{X}_n - \mu)/\sigma$ and $Y = S_n^2/\sigma^2$. It follows from Theorem 8.3.1 that Y and Z are independent, Y has the χ^2 distribution with $n-1$ degrees of freedom, and Z has the standard normal distribution. Finally, define U by

$$U = \frac{Z}{\left(\dfrac{Y}{n-1}\right)^{1/2}}.$$

It follows from the definition of the t distribution that U has the t distribution with $n-1$ degrees of freedom. It is easily seen that U can be rewritten as

$$U = \frac{n^{1/2}(\overline{X}_n - \mu)}{\left(\dfrac{S_n^2}{n-1}\right)^{1/2}}. \tag{8.4.4}$$

The denominator of the expression on the right side of Eq. (8.4.4) is easily recognized as σ' defined in Eq. (8.4.3). ■

The first rigorous proof of Theorem 8.4.2 was given by R. A. Fisher in 1923.

One important aspect of Eq. (8.4.4) is that neither the value of U nor the distribution of U depends on the value of the variance σ^2. In Example 8.4.1, we tried replacing σ in the random variable $Z = n^{1/2}(\overline{X}_n - \mu)/\sigma$ by $\hat{\sigma}$. Instead, Theorem 8.4.2 suggests that we should replace σ by σ' defined in Eq. (8.4.3). If we replace σ by σ', we produce the random variable U in Eq. (8.4.4) that does not involve σ and also has a distribution that does not depend on σ.

The reader should notice that σ' differs from the M.L.E. $\hat{\sigma}$ of σ by a constant factor,

$$\sigma' = \left[\frac{S_n^2}{n-1}\right]^{1/2} = \left(\frac{n}{n-1}\right)^{1/2} \hat{\sigma}. \tag{8.4.5}$$

It can be seen from Eq. (8.4.5) that for large values of n the estimators σ' and $\hat{\sigma}$ will be very close to each other. The estimator σ' will be discussed further in Sec. 8.7.

If the sample size n is large, the probability that the estimator σ' will be close to σ is high. Hence, replacing σ by σ' in the random variable Z will not greatly change the standard normal distribution of Z. For this reason, it is plausible that the t distribution with $n-1$ degrees of freedom should be close to the standard normal distribution if n is large. We shall return to this point more formally later in this section.

Example 8.4.3

Rain from Seeded Clouds. Return to Example 8.4.2. Under the assumption that the observations $X_1, \ldots, X_n$ (log-rainfalls) are independent with common normal distribution, the distribution of $U = n^{1/2}(\overline{X}_n - \mu)/\sigma'$ is the t distribution with $n-1$ degrees of freedom. With $n = 26$, the table of the t distribution tells us that the 0.9 quantile of the t distribution with 25 degrees of freedom is 1.316, so $\Pr(U \le 1.316) = 0.9$. It follows that

$$\Pr\left(\overline{X}_n \le \mu + 0.2581\sigma'\right) = 0.9,$$

because $1.316/(26)^{1/2} = 0.2581$. That is, the probability is 0.9 that $\overline{X}_n$ will be no more than 0.2581 times σ' above μ. Of course, σ' is a random variable as well as $\overline{X}_n$, so this result is not as informative as we might have hoped. In Sections 8.5 and 8.6, we will show how to make use of the t distribution to make some standard inferences about the unknown mean μ. ◀

Relation to the Cauchy Distribution and to the Standard Normal Distribution

It can be seen from Eq. (8.4.2) (and Fig. 8.4) that the p.d.f. $g(x)$ is a symmetric, bell-shaped function with its maximum value at $x = 0$. Thus, its general shape is similar to that of the p.d.f. of a normal distribution with mean 0. However, as $x \to \infty$ or $x \to -\infty$, the tails of the p.d.f. $g(x)$ approach 0 much more slowly than do the tails of the p.d.f. of a normal distribution. In fact, it can be seen from Eq. (8.4.2) that the t distribution with one degree of freedom is the Cauchy distribution, which was defined in Example 4.1.8. The p.d.f. of the Cauchy distribution was sketched in Fig. 4.3. It was shown in Example 4.1.8 that the mean of the Cauchy distribution does not exist, because the integral that specifies the value of the mean is not absolutely convergent. It follows that, although the p.d.f. of the t distribution with one degree of freedom is symmetric with respect to the point $x = 0$, the mean of this distribution does not exist.

It can also be shown from Eq. (8.4.2) that, as $n \to \infty$, the p.d.f. $g(x)$ converges to the p.d.f. $\phi(x)$ of the standard normal distribution for every value of x ($-\infty < x < \infty$). This follows from Theorem 5.3.3 and the following result:

$$\lim_{m\to\infty} \frac{\Gamma\left(m + \frac{1}{2}\right)}{\Gamma(m)m^{1/2}} = 1. \tag{8.4.6}$$

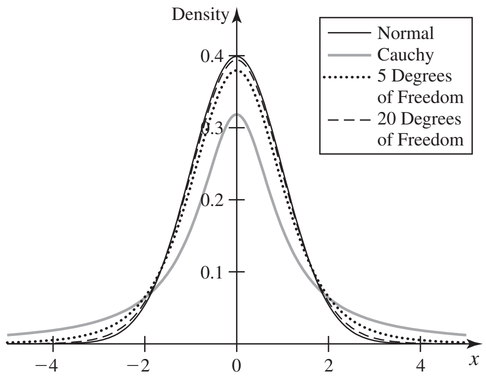


Figure 8.4 p.d.f.'s of standard normal and t distributions.

(See Exercise 7 for a way to prove the above result.) Hence, when n is large, the t distribution with n degrees of freedom can be approximated by the standard normal distribution. Figure 8.4 shows the p.d.f. of the standard normal distribution together with the p.d.f.'s of the t distributions with 1, 5, and 20 degrees of freedom so that the reader can see how the t distributions get closer to normal as the degrees of freedom increase.

A short table of p quantiles for the t distribution with m degrees of freedom for various values of p and m is given at the end of this book. The probabilities in the first line of the table, corresponding to $m = 1$, are those for the Cauchy distribution. The probabilities in the bottom line of the table corresponding to $m = \infty$ are those for the standard normal distribution. Most statistical packages include a function to compute the c.d.f. and the quantile function of an arbitrary t distribution.

Derivation of the p.d.f.

Suppose that the joint distribution of Y and Z is as specified in Definition 8.4.1. Then, because Y and Z are independent, their joint p.d.f. is equal to the product $f_1(y)f_2(z)$, where $f_1(y)$ is the p.d.f. of the χ^2 distribution with m degrees of freedom and $f_2(z)$ is the p.d.f. of the standard normal distribution. Let X be defined by Eq. (8.4.1) and, as a convenient device, let $W = Y$. We shall determine first the joint p.d.f. of X and W.

From the definitions of X and W,

$$Z = X\left(\frac{W}{m}\right)^{1/2} \quad \text{and} \quad Y = W. \tag{8.4.7}$$

The Jacobian of the transformation (8.4.7) from X and W to Y and Z is $(W/m)^{1/2}$. The joint p.d.f. $f(x, w)$ of X and W can be obtained from the joint p.d.f. $f_1(y)f_2(z)$ by replacing y and z by the expressions given in (8.4.7) and then multiplying the result by $(w/m)^{1/2}$. It is then found that the value of $f(x, w)$ is as follows, for $-\infty < x < \infty$ and $w > 0$:

$$\begin{aligned} f(x, w) &= f_1(w) f_2\left(x\left[\frac{w}{m}\right]^{1/2}\right)\left(\frac{w}{m}\right)^{1/2} \\ &= c w^{(m+1)/2-1} \exp\left[-\frac{1}{2}\left(1+\frac{x^2}{m}\right)w\right], \end{aligned} \tag{8.4.8}$$

where

$$c = \left[2^{(m+1)/2}(m\pi)^{1/2}\Gamma\left(\frac{m}{2}\right)\right]^{-1}.$$

The marginal p.d.f. $g(x)$ of X can be obtained from Eq. (8.4.8) by using the relation

$$g(x) = \int f(x, w)\,dw$$
$$= c\int_0^\infty w^{(m+1)/2-1}\exp[-wh(x)]\,dw,$$

where $h(x) = [1 + x^2/m]/2$. It follows from Eq. (5.7.10) that

$$g(x) = c\frac{\Gamma((m+1)/2)}{h(x)^{(m+1)/2}}.$$

Substituting the formula for c into this yields the function in (8.4.2).

Summary

Let $X_1, \dots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\sigma' = \left(\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right)^{1/2}$. Then the distribution of $n^{1/2}(\overline{X}_n - \mu)/\sigma'$ is the t distribution with $n-1$ degrees of freedom.

Exercises

1. Suppose that X has the t distribution with m degrees of freedom $(m > 2)$. Show that $\text{Var}(X) = m/(m-2)$. *Hint:* To evaluate $E(X^2)$, restrict the integral to the positive half of the real line and change the variable from x to

$$y = \frac{\frac{x^2}{m}}{1 + \frac{x^2}{m}}.$$

Compare the integral with the p.d.f. of a beta distribution. Alternatively, use Exercise 21 in Sec. 5.7.

2. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown standard deviation σ, and let $\hat{\mu}$ and $\hat{\sigma}$ denote the M.L.E.'s of μ and σ. For the sample size $n = 17$, find a value of k such that

$$\Pr(\hat{\mu} > \mu + k\hat{\sigma}) = 0.95.$$

3. Suppose that the five random variables $X_1, \dots, X_5$ are i.i.d. and that each has the standard normal distribution. Determine a constant c such that the random variable

$$\frac{c(X_1 + X_2)}{(X_3^2 + X_4^2 + X_5^2)^{1/2}}$$

will have a t distribution.

4. By using the table of the t distribution given in the back of this book, determine the value of the integral

$$\int_{-\infty}^{2.5} \frac{dx}{(12 + x^2)^2}.$$

5. Suppose that the random variables X_1 and X_2 are independent and that each has the normal distribution with mean 0 and variance σ^2. Determine the value of

$$\Pr\left[\frac{(X_1 + X_2)^2}{(X_1 - X_2)^2} < 4\right].$$

Hint:

$$(X_1 - X_2)^2 = 2\left[\left(X_1 - \frac{X_1 + X_2}{2}\right)^2 + \left(X_2 - \frac{X_1 + X_2}{2}\right)^2\right].$$

6. In Example 8.2.3, suppose that we will observe $n = 20$ cheese chunks with lactic acid concentrations $X_1, \dots, X_{20}$. Find a number c so that $\Pr(\overline{X}_{20} \le \mu + c\sigma') = 0.95$.

7. Prove the limit formula Eq. (8.4.6). *Hint:* Use Theorem 5.7.4.

8. Let X have the standard normal distribution, and let Y have the t distribution with five degrees of freedom. Explain why $c = 1.63$ provides the largest value of the difference $\Pr(-c < X < c) - \Pr(-c < Y < c)$. *Hint:* Start by looking at Fig. 8.4.

8.5 Confidence Intervals

Confidence intervals provide a method of adding more information to an estimator $\hat{\theta}$ when we wish to estimate an unknown parameter θ. We can find an interval (A, B) that we think has high probability of containing θ. The length of such an interval gives us an idea of how closely we can estimate θ.

Confidence Intervals for the Mean of a Normal Distribution

Example 8.5.1 Rain from Seeded Clouds. In Example 8.3.2, the average of the $n = 26$ log-rainfalls from the seeded clouds is $\overline{X}_n$. This may be a sensible estimator of the μ, the mean log-rainfall from a seeded cloud, but it doesn't give any idea how much stock we should place in the estimator. The standard deviation of $\overline{X}_n$ is $\sigma/(26)^{1/2}$, and we could estimate σ by an estimator like σ' from Eq. (8.4.3). Is there a sensible way to combine these two estimators into an inference that tells us both what we should estimate for μ and how much confidence we should place in the estimator? ◀

Assume that $X_1, \dots, X_n$, form a random sample from the normal distribution with mean μ and variance σ^2. Construct the estimators $\overline{X}_n$ of μ and σ' of σ . We shall now show how to make use of the random variable

$$U = \frac{n^{1/2}(\overline{X}_n - \mu)}{\sigma'} \tag{8.5.1}$$

from Eq. (8.4.4) to address the question at the end of Example 8.5.1. We know that U has the t distribution with $n-1$ degrees of freedom. Hence, we can calculate the c.d.f. of U and/or quantiles of U using either statistical software or tables such as those in the back of this book. In particular, we can compute $\Pr(-c < U < c)$ for every $c > 0$. The inequalities $-c < U < c$ can be translated into inequalities involving μ by making use of the formula for U in Eq. (8.5.1). Simple algebra shows that $-c < U < c$ is equivalent to

$$\overline{X}_n - \frac{c\sigma'}{n^{1/2}} < \mu < \overline{X}_n + \frac{c\sigma'}{n^{1/2}}. \tag{8.5.2}$$

Whatever probability we can assign to the event $\{-c < U < c\}$ we can also assign to the event that Eq. (8.5.2) holds. For example, if $\Pr(-c < U < c) = \gamma$, then

$$\Pr\left(\overline{X}_n - \frac{c\sigma'}{n^{1/2}} < \mu < \overline{X}_n + \frac{c\sigma'}{n^{1/2}}\right) = \gamma. \tag{8.5.3}$$

One must be careful to understand the probability statement in Eq. (8.5.3) as being a statement about the joint distribution of the random variables $\overline{X}_n$ and σ' for fixed values of μ and σ. That is, it is a statement about the sampling distribution of $\overline{X}_n$ and

σ', and is conditional on μ and σ. In particular, it is *not* a statement about μ even if we treat μ as a random variable.

The most popular version of the calculation above is to choose γ and then figure out what c must be in order to make (8.5.3) true. That is, what value of c makes $\Pr(-c < U < c) = \gamma$? Let T_{n-1} denote the c.d.f. of the t distribution with $n-1$ degrees of freedom. Then

$$\gamma = \Pr(-c < U < c) = T_{n-1}(c) - T_{n-1}(-c).$$

Since the t distributions are symmetric around 0, $T_{n-1}(-c) = 1 - T_{n-1}(c)$, so $\gamma = 2T_{n-1}(c) - 1$ or, equivalently, $c = T_{n-1}^{-1}([1+\gamma]/2)$. That is, c must be the $(1+\gamma)/2$ quantile of the t distribution with $n-1$ degrees of freedom.

Example 8.5.2 **Rain from Seeded Clouds.** In Example 8.3.2, we have $n = 26$. If we want $\gamma = 0.95$ in Eq. (8.5.3), then we need c to be the $1.95/2 = 0.975$ quantile of the t distribution with 25 degrees of freedom. This can be found in the table of t distribution quantiles in the back of the book to be $c = 2.060$. We can plug this value into Eq. (8.5.3) and combine the constants $c/n^{1/2} = 2.060/26^{1/2} = 0.404$. Then Eq. (8.5.3) states that regardless of the unknown values of μ and σ, the probability is 0.95 that the two random variables $A = \overline{X}_n - 0.404\sigma'$ and $B = \overline{X}_n + 0.404\sigma'$ will lie on opposite sides of μ. ◀

The interval (A, B), whose endpoints were computed at the end of Example 8.5.2, is called a *confidence interval*.

Definition 8.5.1 **Confidence Interval.** Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter (or parameter vector) θ. Let $g(\theta)$ be a real-valued function of θ. Let $A \le B$ be two statistics that have the property that for all values of θ,

$$\Pr(A < g(\theta) < B) \ge \gamma. \tag{8.5.4}$$

Then the random interval (A, B) is called a *coefficient γ confidence interval for $g(\theta)$* or a 100γ *percent confidence interval for* $g(\theta)$. If the inequality "$\ge \gamma$" in Eq. (8.5.4) is an equality for all θ, the confidence interval is called *exact*. After the values of the random variables $X_1, \ldots, X_n$ in the random sample have been observed, the values of $A = a$ and $B = b$ are computed, and the interval (a, b) is called the observed value of the confidence interval.

In Example 8.5.2, $\theta = (\mu, \sigma^2)$, and the interval (A, B) found in that example is an exact 95% confidence interval for $g(\theta) = \mu$.

Based on the discussion preceding Definition 8.5.1, we have established the following.

Theorem 8.5.1 **Confidence Interval for the Mean of a Normal Distribution.** Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2. For each $0 < \gamma < 1$, the interval (A, B) with the following endpoints is an exact coefficient γ confidence interval for μ:

$$A = \overline{X}_n - T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma'}{n^{1/2}},$$

$$B = \overline{X}_n + T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma'}{n^{1/2}}.$$ ■

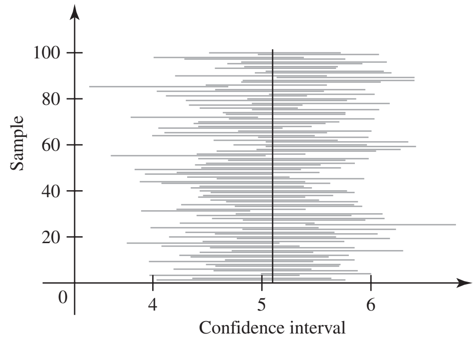


Figure 8.5 A sample of one hundred observed 95% confidence intervals based on samples of size 26 from the normal distribution with mean $\mu = 5.1$ and standard deviation $\sigma = 1.6$. In this figure, 94% of the intervals contain the value of μ.

Example 8.5.3

Rain from Seeded Clouds. In Example 8.5.2, the average of the 26 log-rainfalls from the seeded clouds is $\overline{X}_n = 5.134$. The observed value of σ' is 1.600. The observed values of A and B are, respectively, $a = 5.134 - 0.404 \times 1.600 = 4.488$ and $b = 5.134 + 0.404 \times 1.600 = 5.780$. The observed value of the 95% confidence interval is then (4.488, 5.780). For comparison, the mean unseeded level of 4 is a bit below the lower endpoint of this interval. ◀

Interpretaton of Confidence Intervals The interpretation of the confidence interval (A, B) defined in Definition 8.5.1 is straightforward, so long as one remembers that $\Pr(A < g(\theta) < B) = \gamma$ is a probability statement about the joint distribution of the two random variables A and B given a particular value of θ. Once we compute the observed values a and b, the observed interval (a, b) is not so easy to interpret. For example, some people would like to interpret the interval in Example 8.5.3 as meaning that we are 95% confident that μ is between 4.488 and 5.780. Later in this section, we shall show why such an interpretation is not safe in general. Before observing the data, we can be 95% confident that the random interval (A, B) will contain μ, but after observing the data, the safest interpretation is that (a, b) is simply the observed value of the random interval (A, B). One way to think of the random interval (A, B) is to imagine that the sample that we observed is one of many possible samples that we could have observed (or may yet observe in the future). Each such sample would allow us to compute an observed interval. Prior to observing the samples, we would expect 95% of the intervals to contain μ. Even if we observed many such intervals, we won't know which ones contain μ and which ones don't. Figure 8.5 contains a plot of 100 observed values of confidence intervals, each computed from a sample of size $n = 26$ from the normal distribution with mean $\mu = 5.1$ and standard deviation $\sigma = 1.6$. In this example, 94 of the 100 intervals contain the value of μ.

Example 8.5.4

Acid Concentration in Cheese. In Example 8.2.3, we discussed a random sample of 10 lactic acid measurements from cheese. Suppose that we desire to compute a 90% confidence interval for μ, the unknown mean lactic acid concentration. The number c that we need in Eq. (8.5.3) when $n = 10$ and $\gamma = 0.9$ is the $(1 + 0.9)/2 = 0.95$ quantile of the t distribution with nine degrees of freedom, $c = 1.833$. According to Eq. (8.5.3), the endpoints will be $\overline{X}_n$ plus and minus $1.833\sigma'/(10)^{1/2}$. Suppose that we observe the following 10 lactic acid concentrations as reported by Moore and McCabe (1999, p. D-1):

$$0.86, 1.53, 1.57, 1.81, 0.99, 1.09, 1.29, 1.78, 1.29, 1.58.$$

The average of these 10 values is $\overline{x}_n = 1.379$, and the value of $\sigma' = 0.3277$. The endpoints of the observed value of our 90% confidence interval are then $1.379 - 1.833 \times 0.3277/(10)^{1/2} = 1.189$ and $1.379 + 1.833 \times 0.3277/(10)^{1/2} = 1.569$. ◀

Note: Alternative Definitions of Confidence Interval. Many authors define confidence intervals precisely as we have done here. Some others define the confidence interval to be what we called the observed value of the confidence interval, namely, (a, b), and they need another name for the random interval (A, B). Throughout this book, we shall stay with the definition we have given, but the reader who studies statistics further might encounter the other definition at a later date. Also, some authors define confidence intervals to be closed intervals rather than open intervals.

One-Sided Confidence Intervals

Example 8.5.5 Rain from Seeded Clouds. Suppose that we are interested only in obtaining a lower bound on μ, the mean log-rainfall of seeded clouds. In the spirit of confidence intervals, we could then seek a random variable A such that $\Pr(A < \mu) = \gamma$. If we let $B = \infty$ in Definition 8.5.1, we see that (A, ∞) is then a coefficient γ confidence interval for μ. ◀

For a given confidence coefficient γ, it is possible to construct many different confidence intervals for μ. For example, let $\gamma_2 > \gamma_1$ be two numbers such that $\gamma_2 - \gamma_1 = \gamma$, and let U be as in Eq. (8.5.1). Then

$$\Pr\left(T_{n-1}^{-1}(\gamma_1) < U < T_{n-1}^{-1}(\gamma_2)\right) = \gamma,$$

and the following statistics are the endpoints of a coefficient γ confidence interval for μ:

$$A = \overline{X}_n + T_{n-1}^{-1}(\gamma_1)\frac{\sigma'}{n^{1/2}} \quad \text{and} \quad B = \overline{X}_n + T_{n-1}^{-1}(\gamma_2)\frac{\sigma'}{n^{1/2}}.$$

Among all such coefficient γ confidence intervals, the symmetric interval with $\gamma_1 = 1 - \gamma_2$ is the shortest one.

Nevertheless, there are cases, such as Example 8.5.5, in which an asymmetric confidence interval is useful. In general, it is a simple matter to extend Definition 8.5.1 to allow either $A = -\infty$ or $B = \infty$ so that the confidence interval either has the form $(-\infty, B)$ or (A, ∞).

Definition 8.5.2 One-Sided Confidence Intervals/Limits. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter (or parameter vector) θ. Let $g(\theta)$ be a real-valued function of θ. Let A be a statistic that has the property that for all values of θ,

$$\Pr(A < g(\theta)) \geq \gamma. \tag{8.5.5}$$

Then the random interval (A, ∞) is called a *one-sided coefficient γ confidence interval for $g(\theta)$* or a *one-sided 100γ percent confidence interval for $g(\theta)$*. Also, A is called a *coefficient γ lower confidence limit for $g(\theta)$* or a *100γ percent lower confidence limit for $g(\theta)$*. Similarly, if B is a statistic such that

$$\Pr(g(\theta) < B) \geq \gamma, \tag{8.5.6}$$

then $(-\infty, B)$ is a *one-sided coefficient γ confidence interval for $g(\theta)$* or a *one-sided 100γ percent confidence interval for $g(\theta)$* and B is a *coefficient γ upper confidence limit*

for $g(\theta)$ or a 100γ *percent upper confidence limit for* $g(\theta)$. If the inequality "$\geq \gamma$" in either Eq. (8.5.5) or Eq. (8.5.6) is equality for all θ, the corresponding confidence interval and confidence limit are called *exact*.

The following result follows in much the same way as Theorem 8.5.1.

Theorem 8.5.2 One-Sided Confidence Intervals for the Mean of a Normal Distribution. Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2. For each $0 < \gamma < 1$, the following statistics are, respectively, exact lower and upper coefficient γ confidence limits for μ:

$$A = \overline{X}_n - T_{n-1}^{-1}(\gamma)\,\frac{\sigma'}{n^{1/2}},$$

$$B = \overline{X}_n + T_{n-1}^{-1}(\gamma)\,\frac{\sigma'}{n^{1/2}}.$$ ■

Example 8.5.6 Rain from Seeded Clouds. In Example 8.5.5, suppose that we want a 90% lower confidence limit for μ. We find $T_{25}^{-1}(0.9) = 1.316$. Using the observed data from Example 8.5.3, we compute the observed lower confidence limit as

$$a = 5.134 - 1.316\frac{1.600}{26^{1/2}} = 4.727.$$ ◄

Confidence Intervals for Other Parameters

Example 8.5.7 Lifetimes of Electronic Components. Recall the company in Example 8.1.3 that is estimating the failure rate θ of electronic components based on a sample of $n = 3$ observed lifetimes X_1, X_2, X_3. The statistic $T = \sum_{i=1}^{3} X_i$ was used in Examples 8.1.4 and 8.1.5 to make some inferences. We can use the distribution of T to construct confidence intervals for θ. Recall from Example 8.1.5 that θT has the gamma distribution with parameters 3 and 1 for all θ. Let G stand for the c.d.f. of this gamma distribution. Then $\Pr(\theta T < G^{-1}(\gamma)) = \gamma$ for all θ. It follows that $\Pr(\theta < G^{-1}(\gamma)/T) = \gamma$ for all θ, and $G^{-1}(\gamma)/T$ is an exact coefficient γ upper confidence limit for θ. For example, if the company would like to have a random variable B so that they can be 98% confident that the failure rate θ is bounded above by B, they can find $G^{-1}(0.98) = 7.516$. Then $B = 7.516/T$ is the desired upper confidence limit. ◄

In Example 8.5.7, the random variable θT has the property that its distribution is the same for all θ. The random variable U in Eq. (8.5.1) has the property that its distribution is the same for all μ and σ. Such random variables greatly facilitate the construction of confidence intervals.

Definition 8.5.3 Pivotal. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter (or vector of parameters) θ. Let $V(\boldsymbol{X}, \theta)$ be a random variable whose distribution is the same for all θ. Then V is called a *pivotal quantity* (or simply a *pivotal*).

In order to be able to use a pivotal to construct a confidence interval for $g(\theta)$, one needs to be able to "invert" the pivotal. That is, one needs a function $r(v, \boldsymbol{x})$ such that

$$r\,(V(\boldsymbol{X}, \theta), \boldsymbol{X}) = g(\theta). \tag{8.5.7}$$

If such a function exists, then one can use it to construct confidence intervals.

Theorem 8.5.3 **Confidence Interval from a Pivotal.** Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter (or vector of parameters) θ. Suppose that a pivotal V exists. Let G be the c.d.f. of V, and assume that G is continuous. Assume that a function r exists as in Eq. (8.5.7), and assume that $r(v, \boldsymbol{x})$ is strictly increasing in v for each $\boldsymbol{x}$. Let $0 < \gamma < 1$ and let $\gamma_2 > \gamma_1$ be such that $\gamma_2 - \gamma_1 = \gamma$. Then the following statistics are the endpoints of an exact coefficient γ confidence interval for $g(\theta)$:

$$A = r\left(G^{-1}(\gamma_1), \boldsymbol{X}\right),$$
$$B = r\left(G^{-1}(\gamma_2), \boldsymbol{X}\right).$$

If $r(v, \boldsymbol{x})$ is strictly decreasing in v for each $\boldsymbol{x}$, then switch the definitions of A and B.

Proof If $r(v, \boldsymbol{x})$ is strictly increasing in v for each $\boldsymbol{x}$, we have

$$V(\boldsymbol{X}, \theta) < c \text{ if and only if } g(\theta) < r(c, \boldsymbol{X}). \tag{8.5.8}$$

Let $c = G^{-1}(\gamma_i)$ in Eq. (8.5.8) for each of $i = 1, 2$ to obtain

$$\Pr(g(\theta) < A) = \gamma_1,$$
$$\Pr(g(\theta) < B) = \gamma_2. \tag{8.5.9}$$

Because V has a continuous distribution and r is strictly increasing,

$$\Pr(A = g(\theta)) = \Pr(V(\boldsymbol{X}, \theta) = G^{-1}(\gamma_1)) = 0.$$

Similarly, $\Pr(B = g(\theta)) = 0$. The two equations in (8.5.9) combine to give $\Pr(A < g(\theta) < B) = \gamma$. The proof when r is strictly decreasing is similar and is left to the reader. ■

Example 8.5.8 **Pivotal for Estimating the Variance of a Normal Distribution.** Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2 In Theorem 8.3.1, we found that the random variable $V(\boldsymbol{X}, \theta) = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2/\sigma^2$ has the χ^2 distribution with $n-1$ degrees of freedom for all $\theta = (\mu, \sigma^2)$. This makes V a pivotal. The reader can use this pivotal in Exercise 5 in this section to find a confidence interval of $g(\theta) = \sigma^2$. ◄

Sometimes pivotals do not exist. This is common when the data have a discrete distribution.

Example 8.5.9 **A Clinical Trial.** Consider the imipramine treatment group in the clinical trial in Example 2.1.4. Let θ stand for the proportion of successes among a very large population of imipramine patients. Suppose that the clinicians desire a random variable A such that, for all θ, $\Pr(A < \theta) \geq 0.9$. That is, they want to be 90% confident that the success proportion is at least A. The observable data consist of the number X of successes in a random sample of $n = 40$ patients. No pivotal exists in this example, and confidence intervals are more difficult to construct. In Example 9.1.16, we shall see a method that applies to this case. ◄

Even with discrete data, if the sample size is large enough to apply the central limit theorem, one can find approximate confidence intervals.

Example 8.5.10 **Approximate Confidence Interval for Poisson Mean.** Suppose that $X_1, \ldots, X_n$ have the Poisson distribution with unknown mean θ. Suppose that n is large enough so that

$\overline{X}_n$ has approximately a normal distribution. In Example 6.3.8 on page 365, we found that

$$\Pr\left(|2\overline{X}_n^{1/2} - 2\theta^{1/2}| < c\right) \approx 2\Phi(cn^{1/2}) - 1. \tag{8.5.10}$$

After we observe $\overline{X}_n = x$, Eq. (8.5.10) says that

$$\left(-c + 2x^{1/2}, c + 2x^{1/2}\right) \tag{8.5.11}$$

is the observed value of an approximate confidence interval for $2\theta^{1/2}$ with coefficient $2\Phi(cn^{1/2}) - 1$. For example, if $c = 0.196$ and $n = 100$, then $2\Phi(cn^{1/2}) - 1 = 0.95$. The inverse of $g(\theta) = 2\theta^{1/2}$ is $g^{-1}(y) = y^2/4$, which is an increasing function of y for $y \geq 0$. If both endpoints of (8.5.11) are nonnegative, then we know that $2\theta^{1/2}$ is in the interval (8.5.11) if and only if θ is in the interval

$$\left(\frac{1}{4}[-c + 2x^{1/2}]^2, \frac{1}{4}[c + 2x^{1/2}]^2\right). \tag{8.5.12}$$

If $-c + 2x^{1/2} < 0$, the left endpoints of (8.5.11) and (8.5.12) should be replaced by 0. With this modification, (8.5.12) is the observed value of an approximate coefficient $2\Phi(cn^{1/2}) - 1$ confidence interval for θ. ◀

Shortcoming of Confidence Intervals

Interpretation of Confidence Intervals Let (A, B) be a coefficient γ confidence interval for a parameter θ, and let (a, b) be the observed value of the interval. It is important to understand that it is *not* correct to say that θ lies in the interval (a, b) with *probability* γ. We shall explain this point further here. *Before* the values of the statistics $A(X_1, \ldots, X_n)$ and $B(X_1, \ldots, X_n)$ are observed, these statistics are random variables. It follows, therefore, from Definition 8.5.1 that θ will lie in the random interval having endpoints $A(X_1, \ldots, X_n)$ and $B(X_1, \ldots, X_n)$ with probability γ. *After* the specific values $A(X_1, \ldots, X_n) = a$ and $B(X_1, \ldots, X_n) = b$ have been observed, it is not possible to assign a probability to the event that θ lies in the specific interval (a, b) without regarding θ as a random variable, which itself has a probability distribution. In order to calculate the probability that θ lies in the interval (a, b), it is necessary first to assign a prior distribution to θ and then use the resulting posterior distribution. Instead of assigning a prior distribution to the parameter θ, many statisticians prefer to state that there is *confidence* γ, rather than probability γ, that θ lies in the interval (a, b). Because of this distinction between confidence and probability, the meaning and the relevance of confidence intervals in statistical practice is a somewhat controversial topic.

Information Can Be Ignored In accordance with the preceding explanation, the interpretation of a confidence coefficient γ for a confidence interval is as follows: Before a sample is taken, there is probability γ that the interval that will be constructed from the sample will include the unknown value of θ. After the sample values are observed, however, there might be additional information about whether or not the interval formed from these particular values actually does include θ. How to adjust the confidence coefficient γ in the light of this information is another controversial topic.

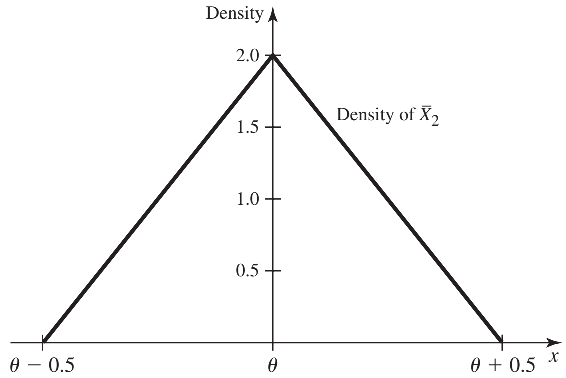


Figure 8.6 p.d.f. of $\overline{X}_2$ in Example 8.5.11.

Example 8.5.11

Uniforms on an Interval of Length One. Suppose that two observations X_1 and X_2 are taken at random from the uniform distribution on the interval $\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$, where the value of θ is unknown $(-\infty < \theta < \infty)$. If we let $Y_1 = \min\{X_1, X_2\}$ and $Y_2 = \max\{X_1, X_2\}$, then

$$\begin{aligned}\Pr(Y_1 < \theta < Y_2) &= \Pr(X_1 < \theta < X_2) + \Pr(X_2 < \theta < X_1)\\ &= \Pr(X_1 < \theta)\Pr(X_2 > \theta) + \Pr(X_2 < \theta)\Pr(X_1 > \theta)\\ &= (1/2)(1/2) + (1/2)(1/2) = 1/2. \end{aligned} \tag{8.5.13}$$

It follows from Eq. (8.5.13) that (Y_1, Y_2) is a confidence interval for θ with confidence coefficient 1/2. However, the analysis can be carried further.

Since both observations X_1 and X_2 must be at least $\theta - (1/2)$, and both must be at most $\theta + (1/2)$, we know with certainty that $Y_1 \ge \theta - (1/2)$ and $Y_2 \le \theta + (1/2)$. In other words, we know with certainty that

$$Y_2 - (1/2) \le \theta \le Y_1 + (1/2). \tag{8.5.14}$$

Suppose now that $Y_1 = y_1$ and $Y_2 = y_2$ are observed such that $(y_2 - y_1) > 1/2$. Then $y_1 < y_2 - (1/2)$, and it follows from Eq. (8.5.14) that $y_1 < \theta$. Moreover, because $y_1 + (1/2) < y_2$, it also follows from Eq. (8.5.14) that $\theta < y_2$. Thus, if $(y_2 - y_1) > 1/2$, then $y_1 < \theta < y_2$. In other words, if $(y_2 - y_1) > 1/2$, then we know with certainty that the observed value (y_1, y_2) of the confidence interval includes the unknown value of θ, even though the confidence coefficient of this interval is only 1/2.

Indeed, even when $(y_2 - y_1) \le 1/2$, the closer the value of $(y_2 - y_1)$ is to 1/2, the more certain we feel that the interval (y_1, y_2) includes θ. Also, the closer the value of $(y_2 - y_1)$ is to 0, the more certain we feel that the interval (y_1, y_2) does not include θ. However, the confidence coefficient necessarily remains 1/2 and does not depend on the observed values y_1 and y_2.

This example also helps to illustrate the statement of caution made at the end of Sec. 8.1. In this problem, it might seem natural to estimate θ by $\overline{X}_2 = 0.5(X_1 + X_2)$. Using the methods of Sec. 3.9, we can find the p.d.f. of $\overline{X}_2$:

$$g(x) = \begin{cases} 4x - 4\theta + 2 & \text{if } \theta - \frac{1}{2} < x \le \theta,\\ 4\theta - 4x + 2 & \text{if } \theta < x < \theta + \frac{1}{2},\\ 0 & \text{otherwise.}\end{cases}$$

Figure 8.6 shows the p.d.f. g, which is triangular. This makes it fairly simple to compute the probability that $\overline{X}_2$ is close to θ:

$$\Pr(|\overline{X}_2 - \theta| < c) = 4c(1 - c),$$

for $0 < c < 1/2$, and the probability is 1 for $c \ge 1/2$. For example, if $c = 0.3$, $\Pr(|\overline{X}_2 - \theta| < 0.3) = 0.84$. However, the random variable $Z = Y_2 - Y_1$ contains useful information that is not accounted for in this calculation. Indeed, the conditional distribution of $\overline{X}_2$ given $Z = z$ is uniform on the interval $\left[\theta - \frac{1}{2}(1-z), \theta + \frac{1}{2}(1-z)\right]$. We see that the larger the observed value of z, the shorter the range of possible values of $\overline{X}_2$. In particular, the conditional probability that $\overline{X}_2$ is close to θ given $Z = z$ is

$$\Pr(|\overline{X}_2 - \theta| < c | Z = z) = \begin{cases} \frac{2c}{1-z} & \text{if } c \le (1-z)/2, \\ 1 & \text{if } c > (1-z)/2. \end{cases} \quad (8.5.15)$$

For example, if $z = 0.1$, then $\Pr(|\overline{X}_2 - \theta| < 0.3 | Z = 0.1) = 0.6667$, which is quite a bit smaller than the marginal probability of 0.84. This illustrates why it is not always safe to assume that our estimate is close to the parameter just because the sampling distribution of the estimator had high probability of being close. There may be other information available that suggests to us that the estimate is not as close as the sampling distribution suggests, or that it is closer than the sampling distribution suggests. (The reader should calculate $\Pr(|\overline{X}_2 - \theta| < 0.3 | Z = 0.9)$ for the other extreme.) ◀

In the next section, we shall discuss Bayesian methods for analyzing a random sample from a normal distribution for which both the mean μ and the variance σ^2 are unknown. We shall assign a joint prior distribution to μ and σ^2, and shall then calculate the posterior probability that μ belongs to any given interval (a, b). It can be shown [see, e.g., DeGroot (1970)] that if the joint prior p.d.f. of μ and σ^2 is fairly smooth and does not assign high probability to any particular small set of values of μ and σ^2, and if the sample size n is large, then the confidence coefficient assigned to a particular confidence interval (A, B) for the mean μ will be approximately equal to the posterior probability that μ lies in the observed interval (a, b). An example of this approximate equality is included in the next section. Therefore, under these conditions, the differences between the results obtained by the practical application of methods based on confidence intervals and methods based on prior probabilities will be small. Nevertheless interpretations of these methods will differ. As an aside, a Bayesian analysis of Example 8.5.11 will necessarily take into account the extra information contained in the random variable Z. See Exercise 10 for an example.

Summary

Let $X_1, \ldots, X_n$ be a random sample of independent random variables from the normal distribution with mean μ and variance σ^2. Let the observed values be $x_1, \ldots, x_n$. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\sigma'^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. The interval $(\overline{X}_n - c\sigma'/n^{1/2}, \overline{X} + c\sigma'/n^{1/2})$ is a coefficient γ confidence interval for μ, where c is the $(1+\gamma)/2$ quantile of the t distribution with $n - 1$ degrees of freedom.

Exercises

1. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance σ^2. Let Φ stand for the c.d.f. of the standard normal distribution, and let Φ^{-1} be its inverse. Show that the following interval is a coefficient γ confidence interval for μ if $\overline{X}_n$ is the observed average of the data values:

$$\left(\overline{X}_n - \Phi^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma}{n^{1/2}}, \overline{X}_n + \Phi^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma}{n^{1/2}}\right).$$

2. Suppose that a random sample of eight observations is taken from the normal distribution with unknown mean μ and unknown variance σ^2, and that the observed values are 3.1, 3.5, 2.6, 3.4, 3.8, 3.0, 2.9, and 2.2. Find the shortest confidence interval for μ with each of the following three confidence coefficients: **(a)** 0.90, **(b)** 0.95, and **(c)** 0.99.

3. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2, and let the random variable L denote the length of the shortest confidence interval for μ that can be constructed from the observed values in the sample. Find the value of $E(L^2)$ for the following values of the sample size n and the confidence coefficient γ:

a. $n = 5, \gamma = 0.95$ **d.** $n = 8, \gamma = 0.90$

b. $n = 10, \gamma = 0.95$ **e.** $n = 8, \gamma = 0.95$

c. $n = 30, \gamma = 0.95$ **f.** $n = 8, \gamma = 0.99$

4. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance σ^2. How large a random sample must be taken in order that there will be a confidence interval for μ with confidence coefficient 0.95 and length less than 0.01σ?

5. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Describe a method for constructing a confidence interval for σ^2 with a specified confidence coefficient γ $(0 < \gamma < 1)$. *Hint:* Determine constants c_1 and c_2 such that

$$\Pr\left[c_1 < \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{\sigma^2} < c_2\right] = \gamma.$$

6. Suppose that $X_1, \ldots, X_n$ form a random sample from the exponential distribution with unknown mean μ. Describe a method for constructing a confidence interval for μ with a specified confidence coefficient γ $(0 < \gamma < 1)$. *Hint:* Determine constants c_1 and c_2 such that $\Pr[c_1 < (1/\mu)\sum_{i=1}^{n} X_i < c_2] = \gamma$.

7. In the June 1986 issue of *Consumer Reports*, some data on the calorie content of beef hot dogs is given. Here are the numbers of calories in 20 different hot dog brands:

186, 181, 176, 149, 184, 190, 158, 139, 175, 148,
152, 111, 141, 153, 190, 157, 131, 149, 135, 132.

Assume that these numbers are the observed values from a random sample of twenty independent normal random variables with mean μ and variance σ^2, both unknown. Find a 90% confidence interval for the mean number of calories μ.

8. At the end of Example 8.5.11, compute the probability that $|\overline{X}_2 - \theta| < 0.3$ given $Z = 0.9$. Why is it so large?

9. In the situation of Example 8.5.11, suppose that we observe $X_1 = 4.7$ and $X_2 = 5.3$.

a. Find the 50% confidence interval described in Example 8.5.11.

b. Find the interval of possible θ values that are consistent with the observed data.

c. Is the 50% confidence interval larger or smaller than the set of possible θ values?

d. Calculate the value of the random variable $Z = Y_2 - Y_1$ as described in Example 8.5.11.

e. Use Eq. (8.5.15) to compute the conditional probability that $|\overline{X}_2 - \theta| < 0.1$ given Z equal to the value computed in part (d).

10. In the situation of Exercise 9, suppose that a prior distribution is used for θ with p.d.f. $\xi(\theta) = 0.1\exp(-0.1\theta)$ for $\theta > 0$. (This is the exponential distribution with parameter 0.1.)

a. Prove that the posterior p.d.f. of θ given the data observed in Exercise 9 is

$$\xi(\theta|\boldsymbol{x}) = \begin{cases} 4.122\exp(-0.1\theta) & \text{if } 4.8 < \theta < 5.2, \\ 0 & \text{otherwise.} \end{cases}$$

b. Calculate the posterior probability that $|\theta - \overline{x}_2| < 0.1$, where $\overline{x}_2$ is the observed average of the data values.

c. Calculate the posterior probability that θ is in the confidence interval found in part (a) of Exercise 9.

d. Can you explain why the answer to part (b) is so close to the answer to part (e) of Exercise 9? *Hint:* Compare the posterior p.d.f. in part (a) to the function in Eq. (8.5.15).

11. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with parameter p. Let $\overline{X}_n$ be the sample average. Use the variance stabilizing transformation found in Exercise 5 of Section 6.5 to construct an approximate coefficient γ confidence interval for p.

12. Complete the proof of Theorem 8.5.3 by dealing with the case in which $r(v, \boldsymbol{x})$ is strictly decreasing in v for each $\boldsymbol{x}$.

★ 8.6 Bayesian Analysis of Samples from a Normal Distribution

When we are interested in constructing a prior distribution for the parameters μ and σ^2 of a normal distribution, it is more convenient to work with $\tau = 1/\sigma^2$, called the precision. *A conjugate family of prior distributions is introduced for μ and τ, and the posterior distribution is derived. Interval estimates of μ can be constructed from the posterior and these are similar to confidence intervals in form, but they are interpreted differently.*

The Precision of a Normal Distribution

Example 8.6.1 Rain from Seeded Clouds. In Example 8.3.1, we mentioned that it was of interest whether the mean log-rainfall μ from seeded clouds exceeded the mean log-rainfall of unseeded clouds, namely, 4. Although we were able to find an estimator of μ and we were able to construct a confidence interval for μ, we have not yet directly addressed the question of whether or not $\mu > 4$ or how likely it is that $\mu > 4$. If we construct a joint prior distribution for both μ and σ^2, we can then find the posterior distribution of μ and finally provide direct answers to these questions. ◀

Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. In this section, we shall consider the assignment of a joint prior distribution to the parameters μ and σ^2 and study the posterior distribution that is then derived from the observed values in the sample. Manipulating prior and posterior distributions for the parameters of a normal distribution turns out to be simpler if we reparameterize from μ and σ^2 to μ and $\tau = 1/\sigma^2$.

Definition 8.6.1 Precision of a Normal Distribution. The *precision* τ of a normal distribution is defined as the reciprocal of the variance; that is, $\tau = 1/\sigma^2$.

If a random variable has the normal distribution with mean μ and precision τ, then its p.d.f. $f(x|\mu, \tau)$ is specified as follows, for $-\infty < x < \infty$:

$$f(x|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left[-\frac{1}{2}\tau(x-\mu)^2\right].$$

Similarly, if $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and precision τ, then their joint p.d.f. $f_n(\boldsymbol{x}|\mu, \tau)$ is as follows, for $-\infty < x_i < \infty$ $(i = 1, \ldots, n)$:

$$f_n(x|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left[-\frac{1}{2}\tau\sum_{i=1}^{n}(x_i-\mu)^2\right].$$

A Conjugate Family of Prior Distributions

We shall now describe a conjugate family of joint prior distributions for μ and τ. We shall specify the joint distribution of μ and τ by specifying both the conditional distribution of μ given τ and the marginal distribution of τ. In particular, we shall assume that the conditional distribution of μ for each given value of τ is a normal distribution for which the precision is proportional to the given value of τ, and also

that the marginal distribution of τ is a gamma distribution. The family of all joint distributions of this type is a conjugate family of joint prior distributions. If the joint prior distribution of μ and τ belongs to this family, then for every possible set of observed values in the random sample, the joint posterior distribution of μ and τ will also belong to the family. This result is established in Theorem 8.6.1. We shall use the following notation in the theorem and the remainder of this section:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_n^2 = \sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

Theorem 8.6.1 Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown precision τ ($-\infty < \mu < \infty$ and $\tau > 0$). Suppose also that the joint prior distribution of μ and τ is as follows: The conditional distribution of μ given τ is the normal distribution with mean μ_0 and precision $\lambda_0\tau$ ($-\infty < \mu_0 < \infty$ and $\lambda_0 > 0$), and the marginal distribution of τ is the gamma distribution with parameters α_0 and β_0 ($\alpha_0 > 0$ and $\beta_0 > 0$). Then the joint posterior distribution of μ and τ, given that $X_i = x_i$ for $i = 1, \ldots, n$, is as follows: The conditional distribution of μ given τ is the normal distribution with mean μ_1 and precision $\lambda_1\tau$, where

$$\mu_1 = \frac{\lambda_0\mu_0 + n\overline{x}_n}{\lambda_0 + n} \quad \text{and} \quad \lambda_1 = \lambda_0 + n, \tag{8.6.1}$$

and the marginal distribution of τ is the gamma distribution with parameters α_1 and β_1, where

$$\alpha_1 = \alpha_0 + \frac{n}{2} \quad \text{and} \quad \beta_1 = \beta_0 + \frac{1}{2}s_n^2 + \frac{n\lambda_0(\overline{x}_n - \mu_0)^2}{2(\lambda_0 + n)}. \tag{8.6.2}$$

Proof The joint prior p.d.f. $\xi(\mu, \tau)$ of μ and τ can be found by multiplying the conditional p.d.f. $\xi_1(\mu|\tau)$ of μ given τ by the marginal p.d.f. $\xi_2(\tau)$ of τ. By the conditions of the theorem, we have, for $-\infty < \mu < \infty$ and $\tau > 0$,

$$\xi_1(\mu|\tau) \propto \tau^{1/2} \exp\left[-\frac{1}{2}\lambda_0\tau(\mu - \mu_0)^2\right]$$

and

$$\xi_2(\tau) \propto \tau^{\alpha_0 - 1}e^{-\beta_0\tau}.$$

A constant factor involving neither μ nor τ has been dropped from the right side of each of these relations.

The joint posterior p.d.f. $\xi(\mu, \tau|\boldsymbol{x})$ for μ and τ satisfies the relation

$$\begin{aligned}\xi(\mu, \tau|\boldsymbol{x}) &\propto f_n(\boldsymbol{x}|\mu, \tau)\xi_1(\mu|\tau)\xi_2(\tau) &(8.6.3)\\ &\propto \tau^{\alpha_0+(n+1)/2-1} \exp\left[-\frac{\tau}{2}\left(\lambda_0[\mu - \mu_0]^2 + \sum_{i=1}^{n}(x_i - \mu)^2\right) - \beta_0\tau\right].\end{aligned}$$

By adding and subtracting $\overline{x}_n$ inside the $(x_i - \mu)^2$ terms, we can prove that

$$\sum_{i=1}^{n}(x_i - \mu)^2 = s_n^2 + n(\overline{x}_n - \mu)^2. \tag{8.6.4}$$

Next, combine the last term in Eq. (8.6.4) with the term $\lambda_0(\mu - \mu_0)^2$ in (8.6.3) by completing the square (see Exercise 24 in Sec. 5.6) to get

$$n(\overline{x}_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2 = (\lambda_0 + n)(\mu - \mu_1)^2 + \frac{n\lambda_0(\overline{x}_n - \mu_0)^2}{\lambda_0 + n}, \tag{8.6.5}$$

where μ_1 is defined in Eq. (8.6.1). Combining (8.6.4) with (8.6.5) yields

$$\sum_{i=1}^{n}(x_i-\mu)^2+\lambda_0(\mu-\mu_0)^2=(\lambda_0+n)(\mu-\mu_1)^2+s_n^2 + \frac{n\lambda_0(\overline{x}_n-\mu_0)^2}{\lambda_0+n}. \tag{8.6.6}$$

Using (8.6.2) and $\lambda_1=\lambda_0+n$ together with (8.6.6) allows us to write Eq. (8.6.3) in the form

$$\xi(\mu,\tau|\boldsymbol{x})\propto\left\{\tau^{1/2}\exp\left[-\frac{1}{2}\lambda_1\tau(\mu-\mu_1)^2\right]\right\}(\tau^{\alpha_1-1}e^{-\beta_1\tau}), \tag{8.6.7}$$

where λ_1, α_1, and β_1 are defined by Eqs. (8.6.1) and (8.6.2).

When the expression inside the braces on the right side of Eq. (8.6.7) is regarded as a function of μ for a fixed value of τ, this expression can be recognized as being (except for a factor that depends on neither μ nor τ) the p.d.f. of the normal distribution with mean μ_1 and precision $\lambda_1\tau$. Since the variable μ does not appear elsewhere on the right side of Eq. (8.6.7), it follows that this p.d.f. must be the conditional posterior p.d.f. of μ given τ. It now follows in turn that the expression outside the braces on the right side of Eq. (8.6.7) must be proportional to the marginal posterior p.d.f. of τ. This expression can be recognized as being (except for a constant factor) the p.d.f. of the gamma distribution with parameters α_1 and β_1. Hence, the joint posterior distribution of μ and τ is as specified in the theorem. ■

We shall give a name to the family of joint distributions described in Theorem 8.6.1.

Definition 8.6.2 Normal-Gamma Family of Distributions. Let μ and τ be random variables. Suppose that the conditional distribution of μ given τ is the normal distribution with mean μ_0 and precision $\lambda_0\tau$. Suppose also that the marginal distribution of τ is the gamma distribution with parameters α_0 and β_0. Then we say that the joint distribution of μ and τ is the *normal-gamma distribution with hyperparameters* μ_0, λ_0, α_0, and β_0.

The prior distribution in Theorem 8.6.1 is the normal-gamma distribution with hyperparameters μ_0, λ_0, α_0, and β_0. The posterior distribution derived in that theorem is the normal-gamma distribution with hyperparameters μ_1, λ_1, α_1, and β_1. As in Sec. 7.3, we shall refer to the hyperparameters of the prior distribution as *prior hyperparameters*, and we shall refer to the hyperparameters of the posterior distribution as *posterior hyperparameters*.

By choosing appropriate values of the prior hyperparameters, it is usually possible in a particular problem to find a normal-gamma distribution that approximates an experimenter's actual prior distribution of μ and τ sufficiently well. It should be emphasized, however, that if the joint distribution of μ and τ is a normal-gamma distribution, then μ and τ are not independent. Thus, it is not possible to use a normal-gamma distribution as a joint prior distribution of μ and τ in a problem in which the experimenter wishes μ and τ to be independent in the prior. Although this characteristic of the family of normal-gamma distributions is a deficiency, it is not an important deficiency, because of the following fact: Even if a joint prior distribution under which μ and τ are independent is chosen from outside the conjugate family, it will be found that after just a single value of X has been observed, μ and τ will have a posterior distribution under which they are dependent. In other words, it is not possible for μ

and τ to remain independent in the light of even one observation from the underlying normal distribution.

Example 8.6.2 Acid Concentration in Cheese. Consider again the example of lactic acid concentration in cheese as discussed in Example 8.5.4. Suppose that the concentrations are independent normal random variables with mean μ and precision τ. Suppose that the prior opinion of the experimenters could be expressed as a normal-gamma distribution with hyperparameters $\mu_0 = 1$, $\lambda_0 = 1$, $\alpha_0 = 0.5$, and $\beta_0 = 0.5$. We can use the data on page 487 to find the posterior distribution of μ and τ. In this case, $n = 10$, $\overline{x}_n = 1.379$, and $s_n^2 = 0.9663$. Applying the formulas in Theorem 8.6.1, we get

$$\mu_1 = \frac{1 \times 1 + 10 \times 1.379}{1 + 10} = 1.345, \qquad \lambda_1 = 1 + 10 = 11, \qquad \alpha_1 = 0.5 + \frac{10}{2} = 5.5,$$

$$\beta_1 = 0.5 + \frac{1}{2}0.9663 + \frac{10 \times 1 \times (1.379 - 1)^2}{2(1 + 10)} = 1.0484.$$

So, the posterior distribution of μ and τ is the normal-gamma distribution with these four hyperparameters. In particular, we can now address the issue of variation in lactic acid concentration more directly. For example, we can compute the posterior probability that $\sigma = \tau^{-1/2}$ is larger than some value such as 0.3:

$$\Pr(\sigma > 0.3|\boldsymbol{x}) = \Pr(\tau < 11.11|\boldsymbol{x}) = 0.984.$$

This can be found using any computer program that calculates the c.d.f. of a gamma distribution. Alternatively, we can use the relationship between the gamma and χ^2 distributions that allows us to say that the posterior distribution of $U = 2 \times 1.0484 \times \tau$ is the χ^2 distribution with $2 \times 5.5 = 11$ degrees of freedom. (See Exercise 1 in Sec. 5.7.) Then $\Pr(\tau < 11.11|\boldsymbol{x}) = \Pr(U \leq 23.30|\boldsymbol{x}) \approx 0.982$ by interpolating in the table of the χ^2 distributions in the back of the book. If $\sigma > 0.3$ is considered a large standard deviation, the cheese manufacturer might wish to look into better quality-control measures. ◀

The Marginal Distribution of the Mean

When the joint distribution of μ and τ is a normal-gamma distribution of the type described in Theorem 8.6.1, then the conditional distribution of μ for a given value of τ is a normal distribution and the marginal distribution of τ is a gamma distribution. It is not clear from this specification, however, what the marginal distribution of μ will be. We shall now derive this marginal distribution.

Theorem 8.6.2 Marginal Distribution of the Mean. Suppose that the prior distribution of μ and τ is the normal-gamma distribution with hyperparameters μ_0, λ_0, α_0, and β_0. Then the marginal distribution of μ is related to a t distribution in the following way:

$$\left(\frac{\lambda_0 \alpha_0}{\beta_0}\right)^{1/2} (\mu - \mu_0)$$

has the t distribution with $2\alpha_0$ degrees of freedom.

Proof Since the conditional distribution of μ given τ is the normal distribution with mean μ_0 and variance $(\lambda_0 \tau)^{-1}$, we can use Theorem 5.6.4 to conclude that the conditional distribution of $Z = (\lambda_0 \tau)^{1/2}(\mu - \mu_0)$ given τ is the standard normal distribution. We shall continue to let $\xi_2(\tau)$ be the marginal p.d.f. of τ, and let $\xi_1(\mu|\tau)$

be the conditional p.d.f. of μ given τ. Then the joint p.d.f. of Z and τ is

$$f(z, \tau) = (\lambda_0\tau)^{-1/2}\xi_1((\lambda_0\tau)^{-1/2}z + \mu_0|\tau)\xi_2(\tau) = \phi(z)\xi_2(\tau), \tag{8.6.8}$$

where ϕ is the standard normal p.d.f. of Eq. (5.6.6). We see from Eq. (8.6.8) that Z and τ are independent with Z having the standard normal distribution. Next, let $Y = 2\beta_0\tau$. Using the result of Exercise 1 in Sec. 5.7, we find that the distribution of Y is the gamma distribution with parameters α_0 and 1/2, which is also known as the χ^2 distribution with $2\alpha_0$ degrees of freedom. In summary, Y and Z are independent with Z having the standard normal distribution and Y having the χ^2 distribution with $2\alpha_0$ degrees of freedom. It follows from the definition of the t distributions in Sec. 8.4 that

$$U = \frac{Z}{\left(\frac{Y}{2\alpha_0}\right)^{1/2}} = \frac{(\lambda_0\tau)^{1/2}(\mu - \mu_0)}{\left(\frac{2\beta_0\tau}{2\alpha_0}\right)^{1/2}} = \left(\frac{\lambda_0\alpha_0}{\beta_0}\right)^{1/2}(\mu - \mu_0) \tag{8.6.9}$$

has the t distribution with $2\alpha_0$ degrees of freedom. ■

Theorem 8.6.2 can also be used to find the posterior distribution of μ after data are observed. To do that, just replace μ_0 by μ_1, λ_0 by λ_1, α_0 by α_1, and β_0 by β_1 in the statement of the theorem. The reason for this is that the prior and posterior distributions both have the same form, and the theorem depends only on that form. This same reasoning applies to the discussion that follows, including Theorem 8.6.3.

An alternative way to describe the marginal distribution of μ starts by rewriting (8.6.9) as

$$\mu = \left(\frac{\beta_0}{\lambda_0\alpha_0}\right)^{1/2} U + \mu_0. \tag{8.6.10}$$

Now we see that the distribution of μ can be obtained from a t distribution by translating the t distribution so that it is centered at μ_0 rather than at 0, and also changing the scale factor. This makes it straightforward to find the moments (if they exist) of the distribution of μ.

Theorem 8.6.3 Suppose that μ and τ have the joint normal-gamma distribution with hyperparameters μ_0, λ_0, α_0, and β_0. If $\alpha_0 > 1/2$, then $E(\mu) = \mu_0$. If $\alpha_0 > 1$, then

$$\text{Var}(\mu) = \frac{\beta_0}{\lambda_0(\alpha_0 - 1)}. \tag{8.6.11}$$

Proof The mean and the variance of the marginal distribution of μ can easily be obtained from the mean and the variance of the t distributions that are given in Sec. 8.4. Since U in Eq. (8.6.9) has the t distribution with $2\alpha_0$ degrees of freedom, it follows from Section 8.4 that $E(U) = 0$ if $\alpha_0 > 1/2$ and that $\text{Var}(U) = \alpha_0/(\alpha_0 - 1)$ if $\alpha_0 > 1$. Now use Eq. (8.6.10) to see that if $\alpha_0 > 1/2$, then $E(\mu) = \mu_0$. Also, if $\alpha_0 > 1$, then

$$\text{Var}(\mu) = \left(\frac{\beta_0}{\lambda_0\alpha_0}\right)\text{Var}(U).$$

Eq. (8.6.11) now follows directly. ■

Furthermore, the probability that μ lies in any specified interval can, in principle, be obtained from a table of the t distribution or appropriate software. Most statistical packages include functions that can compute the c.d.f. and the quantile function of

a t distribution with arbitrary degrees of freedom, not just integers. Tables typically deal solely with integer degrees of freedom. If necessary, one can interpolate between adjacent degrees of freedom.

As we pointed out already, we can change the prior hyperparameters to posterior hyperparameters in Theorems 8.6.2 and 8.6.3 and translate them into results concerning the posterior marginal distribution of μ. In particular, the posterior distribution of the following random variable is the t distribution with $2\alpha_1$ degrees of freedom:

$$\left(\frac{\lambda_1\alpha_1}{\beta_1}\right)^{1/2}(\mu - \mu_1). \tag{8.6.12}$$

A Numerical Example

Example 8.6.3

Nursing Homes in New Mexico. In 1988, the New Mexico Department of Health and Social Services recorded information from many of its licensed nursing homes. The data were analyzed by Smith, Piland, and Fisher (1992). In this example, we shall consider the annual medical in-patient days X (measured in hundreds) for a sample of 18 nonrural nursing homes. Prior to observing the data, we shall model the value of X for each nursing home as a normal random variable with mean μ and precision τ. To choose a prior mean and variance for μ and τ, we could speak with experts in the field, but for simplicity, we shall just base these on some additional information we have about the numbers of beds in these nursing homes. There are, on average, 111 beds with a sample standard deviation of 43.5 beds. Suppose that our prior opinion is that there is a 50 percent occupancy rate. Then we can naïvely scale up the mean and standard deviation by a factor of 0.5×365 to obtain a prior mean and standard deviation for the number of in-patient days in a year. In units of hundreds of in-patient days per year, this gives us a mean of $0.5 \times 365 \times 1.11 \approx 200$ and a standard deviation of $0.5 \times 365 \times 0.435 \approx 6300^{1/2}$. To map these values into prior hyperparameters, we shall split the variance of 6300 so that half of it is due to variance between the nursing homes and half is the variance of μ. That is, we shall set $\text{Var}(\mu) = 3150$ and $E(\tau) = 1/3150$. We choose $\alpha_0 = 2$ to reflect only a small amount of prior information. Then, since $E(\tau) = \alpha_0/\beta_0$, we find that $\beta_0 = 6300$. Using $E(\mu) = \mu_0$ and (8.6.11), we get $\mu_0 = 200$ and $\lambda_0 = 2$.

Next, we shall determine an interval for μ centered at the point $\mu_0 = 200$ such that the probability that μ lies in this interval is 0.95. Since the random variable U defined by Eq. (8.6.9) has the t distribution with $2\alpha_0$ degrees of freedom, it follows that, for the numerical values just obtained, the random variable $0.025(\mu - 200)$ has the t distribution with four degrees of freedom. The table of the t distribution gives the 0.975 quantile of the t distribution with four degrees of freedom as 2.776. So,

$$\Pr[-2.776 < 0.025(\mu - 200) < 2.776] = 0.95. \tag{8.6.13}$$

An equivalent statement is that

$$\Pr(89 < \mu < 311) = 0.95. \tag{8.6.14}$$

Thus, under the prior distribution assigned to μ and τ, there is probability 0.95 that μ lies in the interval (89, 311).

Suppose now that the following is our sample of 18 observed numbers of medical in-patient days (in hundreds):

128 281 291 238 155 148 154 232 316 96 146 151 100 213 208 157 48 217.

For these observations, which we denote $\boldsymbol{x}$, $\overline{x}_n = 182.17$ and $s_n^2 = 88678.5$. Then, it follows from Theorem 8.6.1 that the joint posterior distribution of μ and τ is the normal-gamma distribution with hyperparameters

$$\mu_1 = 183.95, \qquad \lambda_1 = 20, \qquad \alpha_1 = 11, \qquad \beta_1 = 50925.37. \tag{8.6.15}$$

Hence, the values of the means and the variances of μ and τ, as found from this joint posterior distribution, are

$$\begin{aligned} E(\mu|\boldsymbol{x}) &= \mu_1 = 183.95, & \mathrm{Var}(\mu|\boldsymbol{x}) &= \frac{\beta_1}{\lambda_1(\alpha_1 - 1)} = 254.63, \\ E(\tau|\boldsymbol{x}) &= \frac{\alpha_1}{\beta_1} = 2.16 \times 10^{-4}, & \mathrm{Var}(\tau|\boldsymbol{x}) &= \frac{\alpha_1}{\beta_1^2} = 4.24 \times 10^{-9}. \end{aligned} \tag{8.6.16}$$

It follows from Eq. (8.6.1) that the mean μ_1 of the posterior distribution of μ is a weighted average of μ_0 and $\overline{x}_n$. In this numerical example, it is seen that μ_1 is quite close to $\overline{x}_n$.

Next, we shall determine the marginal posterior distribution of μ. Let U be the random variable in Eq. (8.6.12), and use the values computed in (8.6.15). Then $U = (0.0657)(\mu - 183.95)$, and the posterior distribution of U is the t distribution with $2\alpha_1 = 22$ degrees of freedom. The 0.975 quantile of this t distribution is 2.074, so

$$\Pr(-2.074 < U < 2.074|\boldsymbol{x}) = 0.95. \tag{8.6.17}$$

An equivalent statement is that

$$\Pr(152.38 < \mu < 215.52|\boldsymbol{x}) = 0.95. \tag{8.6.18}$$

In other words, under the posterior distribution of μ and τ, the probability that μ lies in the interval (152.38, 215.52) is 0.95.

It should be noted that the interval in Eq. (8.6.18) determined from the posterior distribution of μ is much shorter than the interval in Eq. (8.6.14) determined from the prior distribution. This result reflects the fact that the posterior distribution of μ is much more concentrated around its mean than was the prior distribution. The variance of the prior distribution of μ was 3150, and the variance of the posterior distribution is 254.63. Graphs of the prior and posterior p.d.f.'s of μ are in Fig. 8.7 together with the posterior interval (8.6.18). ◀

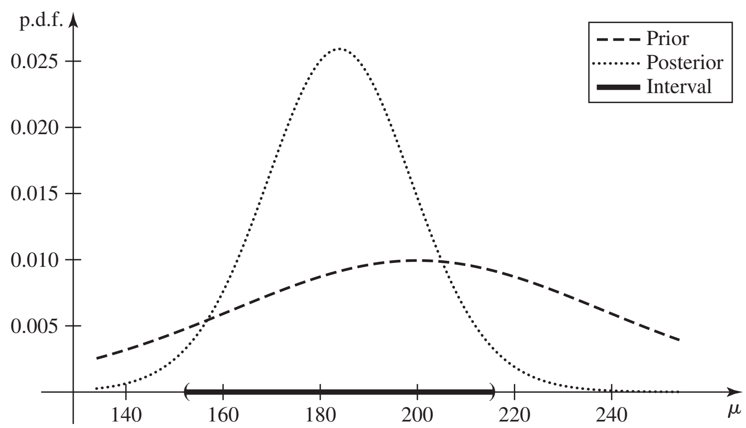


Figure 8.7 Plots of prior and posterior p.d.f.'s of μ in Example 8.6.3. The posterior probability interval (8.6.18) is indicated at the bottom of the graph. The corresponding prior probability interval (8.6.14) would extend far beyond both sides of the plot.

Comparison with Confidence Intervals Continue using the nursing home data from Example 8.6.3. We shall now construct a confidence interval for μ with confidence coefficient 0.95 and compare this interval with the interval in Eq. (8.6.18) for which the posterior probability is 0.95. Since the sample size n in Example 8.6.3 is 18, the random variable U defined by Eq. (8.4.4) on page 481 has the t distribution with 17 degrees of freedom. The 0.975 quantile of this t distribution is 2.110. It now follows from Theorem 8.5.1 that the endpoints of a confidence interval for μ with confidence coefficient 0.95 will be

$$A = \overline{X}_n - 2.110\frac{\sigma'}{n^{1/2}},$$

$$B = \overline{X}_n + 2.110\frac{\sigma'}{n^{1/2}}.$$

When the observed values of $\overline{x}_n = 182.17$ and $s_n^2 = 88678.5$ are used here, we get $\sigma' = (88678.5/17)^{1/2} = 72.22$. The observed confidence interval for μ is then (146.25, 218.09).

This interval is close to the interval (152.38, 215.52) in Eq. (8.6.18), for which the posterior probability is 0.95. The similarity of the two intervals illustrates the statement made at the end of Sec. 8.5. That is, in many problems involving the normal distribution, the method of confidence intervals and the method of using posterior probabilities yield similar results, even though the interpretations of the two methods are quite different.

Improper Prior Distributions

As we discussed at the end of Sec. 7.3 on page 402, it is often convenient to use improper priors that are not real distributions, but do lead to posteriors that are real distributions. These improper priors are chosen more for convenience than to represent anyone's beliefs. When there is a sizeable amount of data, the posterior distribution that results from use of an improper prior is often very close to one that would result from a proper prior distribution. For the case that we have been considering in this section, we can combine the improper prior that we introduced for a location parameter like μ together with the improper prior for a scale parameter like $\sigma = \tau^{-1/2}$ into the usual improper prior for μ and τ. The typical improper prior "p.d.f." for a location parameter was found (in Example 7.3.15) to be the constant function $\xi_1(\mu) = 1$. The typical improper prior "p.d.f." for a scale parameter σ is $g(\sigma) = 1/\sigma$. Since $\sigma = \tau^{-1/2}$, we can apply the techniques of Sec. 3.8 to find the improper "p.d.f." of $\tau = \sigma^{-2}$. The derivative of the inverse function is $-\frac{1}{2}\tau^{-3/2}$, so the improper "p.d.f." of τ would be

$$\left|\frac{1}{2}\tau^{-3/2}\right| g(1/\tau^{-1/2}) = \frac{1}{2}\tau^{-1},$$

for $\tau > 0$. Since this function has infinite integral, we shall drop the factor 1/2 and set $\xi_2(\tau) = \tau^{-1}$. If we act as if μ and τ were independent, then the joint improper prior "p.d.f." for μ and τ is

$$\xi(\mu, \tau) = \frac{1}{\tau}, \quad \text{for } -\infty < \mu < \infty, \tau > 0.$$

If we were to pretend as if this function were a p.d.f., the posterior p.d.f. $\xi(\mu, \tau|\boldsymbol{x})$ would be proportional to

$$\xi(\mu, \tau) f_n(\boldsymbol{x}|\mu, \tau) \propto \tau^{-1}\tau^{n/2} \exp\left(-\frac{\tau}{2}s_n^2 - \frac{n\tau}{2}(\mu - \overline{x}_n)^2\right) \quad (8.6.19)$$
$$= \left\{\tau^{1/2} \exp\left[-\frac{n\tau}{2}(\mu - \overline{x}_n)^2\right]\right\} \tau^{(n-1)/2-1} \exp\left[-\tau\frac{s_n^2}{2}\right].$$

When the expression inside the braces on the far right side of (8.6.19) is regarded as a function of μ for fixed value of τ, this expression can be recognized as being (except for a factor that depends on neither μ nor τ) the p.d.f. of the normal distribution with mean $\overline{x}_n$ and precision $n\tau$. Since the variable μ does not appear elsewhere, it follows that this p.d.f. must be the conditional posterior p.d.f. of μ given τ. It now follows in turn that the expression outside the braces on the far right side of (8.6.19) must be proportional to the marginal posterior p.d.f. of τ. This expression can be recognized as being (except for a constant factor) the p.d.f. of the gamma distribution with parameters $(n-1)/2$ and $s_n^2/2$. This joint distribution would be in precisely the same form as the distribution in Theorem 8.6.1 if our prior distribution had been of the normal-gamma form with hyperparameters $\mu_0 = \beta_0 = \lambda_0 = 0$ and $\alpha_0 = -1/2$. That is, if we pretend as if $\mu_0 = \beta_0 = \lambda_0 = 0$ and $\alpha_0 = -1/2$, and then we apply Theorem 8.6.1, we get the posterior hyperparameters $\mu_1 = \overline{x}_n$, $\lambda_1 = n$, $\alpha_1 = (n-1)/2$, and $\beta_1 = s_n^2/2$.

There is no probability distribution in the normal-gamma family with $\mu_0 = \beta_0 = \lambda_0 = 0$ and $\alpha_0 = -1/2$; however, if we pretend as if this were our prior, then we are said to be using the *usual improper prior distribution*. Notice that the posterior distribution of μ and τ is a real member of the normal-gamma family so long as $n \geq 2$.

Example 8.6.4 An Improper Prior for Seeded Cloud Rainfall. Suppose that we use the usual improper prior for the parameters in Examples 8.3.2 and 8.5.3 with prior hyperparameters $\mu_0 = \beta_0 = \lambda_0 = 0$ and $\alpha_0 = -1/2$. The data summaries are $\bar{x}_n = 5.134$ and $s_n^2 = 63.96$. The posterior distribution will then be the normal-gamma distribution with hyperparameters $\mu_1 = \overline{x}_n = 5.134$, $\lambda_1 = n = 26$, $\alpha_1 = (n-1)/2 = 12.5$, and $\beta_1 = s_n^2/2 = 31.98$. Also, the marginal posterior distribution of μ is given by (7.6.12). In particular,

$$U = \left(\frac{26 \times 12.5}{31.98}\right)^{1/2} (\mu - 5.134) = 3.188(\mu - 5.134) \quad (8.6.20)$$

has the t distribution with 25 degrees of freedom. Suppose that we want an interval (a, b) such that the posterior probability of $a < \mu < b$ is 0.95. The 0.975 quantile of the t distribution with 25 degrees of freedom is 2.060. So, we have that $\Pr(-2.060 < U < 2.060) = 0.95$. Combining this with (8.6.20), we get

$$\Pr(5.134 - 2.060/3.188 < \mu < 5.134 + 2.060/3.188|\boldsymbol{x}) = 0.95.$$

The interval we need runs from $a = 5.134 - 2.060/3.188 = 4.488$ to $b = 5.134 + 2.060/3.188 = 5.780$. Notice that the interval (4.488, 5.780) is precisely the same as the 95% confidence interval for μ that was computed in Example 8.5.3.

Another calculation that we can do with this posterior distribution is to see how likely it is that $\mu > 4$, where 4 is the mean of log-rainfall for unseeded clouds:

$$\Pr(\mu > 4|\boldsymbol{x}) = \Pr(U > 3.188(4 - 5.134)|\boldsymbol{x}) = 1 - T_{25}(-3.615) = 0.9993,$$

where the final value is calculated using statistical software that includes the c.d.f.'s of all t distributions. It appears quite likely, after observing the data, that the mean log-rainfall of seeded clouds is more than 4. ◀

Note: Improper Priors Lead to Confidence Intervals. Example 8.6.4 illustrates one of the more interesting properties of the usual improper prior. If one uses the usual

improper prior with normal data, then the posterior probability is γ that μ is in the observed value of a coefficient γ confidence interval. In general, if we apply (8.6.9) after using an improper prior, we find that the posterior distribution of

$$U = \left(\frac{n(n-1)}{s_n^2}\right)^{1/2} (\mu - \overline{x}_n) \tag{8.6.21}$$

is the t distribution with $n-1$ degrees of freedom. It follows that if $\Pr(-c < U < c) = \gamma$, then

$$\Pr\left(\overline{x}_n - c\frac{\sigma'}{n^{1/2}} < \mu < \overline{x}_n + c\frac{\sigma'}{n^{1/2}} \middle| \boldsymbol{x} \right) = \gamma. \tag{8.6.22}$$

The reader will notice the striking similarity between (8.6.22) and (8.5.3). The difference between the two is that (8.6.22) is a statement about the posterior distribution of μ *after* observing the data, while (8.5.3) is a statement about the conditional distribution of the random variables $\overline{X}_n$ and σ' given μ and σ *before* observing the data. That these two probabilities are the same for all possible data and all possible values of γ follows from the fact that they are both equal to $\Pr(-c < U < c)$ where U is defined either in Eq. (8.4.4) or Eq. (8.6.21). The sampling distribution (conditional on μ and τ) of U is the t distribution with $n-1$ degrees of freedom, as we found in Eq. (8.4.4). The posterior distribution from the improper prior (conditional on the data) of U is also the t distribution with $n-1$ degrees of freedom.

The same kind of thing happens when we try to estimate $\sigma^2 = 1/\tau$. The sampling distribution (conditional on μ and τ) of $V = (n-1)\sigma'^2\tau = (n-1)\sigma'^2/\sigma^2$ is the χ^2 distribution with $n-1$ degrees of freedom, as we saw in Eq. (8.3.11). The posterior distribution from the improper prior (conditional on the data) of V is also the χ^2 distribution with $n-1$ degrees of freedom (see Exercise 4). Therefore, a coefficient γ confidence interval (a, b) for σ^2 based on the sampling distribution of V will satisfy $\Pr(a < \sigma^2 < b|\boldsymbol{x}) = \gamma$ as a posterior probability statement given the data if we used an improper prior.

There are many situations in which the sampling distribution of a pivotal quantity like U above is the same as its posterior distribution when an improper prior is used. A very mathematical treatment of these situations can be found in Schervish (1995, chapter 6). The most common situations are those involving location parameters (like μ) and/or scale parameters (like σ).

Summary

We introduced a family of conjugate prior distributions for the parameters μ and $\tau = 1/\sigma^2$ of a normal distribution. The conditional distribution of μ given τ is normal with mean μ_0 and precision $\lambda_0\tau$, and the marginal distribution of τ is the gamma distribution with parameters α_0 and β_0. If $X_1 = x_1, \ldots, X_n = x_n$ is an observed sample of size n from the normal distribution with mean μ and precision τ, then the posterior distribution of μ given τ is the normal distribution with mean μ_1 and precision $\lambda_1\tau$, and the posterior distribution of τ is the gamma distribution with parameters α_1 and β_1 where the values of μ_1, λ_1, α_1, and β_1 are given in Eq. (8.6.1) and (8.6.2). The marginal posterior distribution of μ is given by saying that $(\lambda_1\alpha_1/\beta_1)^{1/2}(\mu - \mu_1)$ has the t distribution with $2\alpha_1$ degrees of freedom. An interval containing probability $1-\alpha$ of the posterior distribution of μ is

$$\left(\mu_1 - T_{2\alpha_1}^{-1}(1-\alpha/2)\left[\frac{\beta_1}{\alpha_1\lambda_1}\right]^{1/2}, \mu_1 + T_{2\alpha_1}^{-1}(1-\alpha/2)\left[\frac{\beta_1}{\alpha_1\lambda_1}\right]^{1/2}\right).$$

If we use the improper prior with prior hyperparameters $\alpha_0 = -1/2$ and $\mu_0 = \lambda_0 = \beta_0 = 0$, then the random variable $n^{1/2}(\overline{X}_n - \mu)/\sigma'$ has the t distribution with $n-1$ degrees of freedom both as its posterior distribution given the data and as its sampling distribution given μ and σ. Also, $(n-1)\sigma'^2/\sigma^2$ has the χ^2 distribution with $n-1$ degrees of freedom both as its posterior distribution given the data and as its sampling distribution given μ and σ. Hence, if we use the improper prior, interval estimates of μ or σ based on the posterior distribution will also be confidence intervals, and vice versa.

Exercises

1. Suppose that a random variable X has the normal distribution with mean μ and precision τ. Show that the random variable $Y = aX + b$ ($a \neq 0$) has the normal distribution with mean $a\mu + b$ and precision τ/a^2.

2. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ ($-\infty < \mu < \infty$) and known precision τ. Suppose also that the prior distribution of μ is the normal distribution with mean μ_0 and precision λ_0. Show that the posterior distribution of μ, given that $X_i = x_i$ ($i = 1, \dots, n$) is the normal distribution with mean

$$\frac{\lambda_0\mu_0 + n\tau\overline{x}_n}{\lambda_0 + n\tau}$$

and precision $\lambda_0 + n\tau$.

3. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with known mean μ and unknown precision τ ($\tau > 0$). Suppose also that the prior distribution of τ is the gamma distribution with parameters α_0 and β_0 ($\alpha_0 > 0$ and $\beta_0 > 0$). Show that the posterior distribution of τ given that $X_i = x_i$ ($i = 1, \dots, n$) is the gamma distribution with parameters $\alpha_0 + (n/2)$ and

$$\beta_0 + \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2.$$

4. Suppose that $X_1, \dots, X_n$ are i.i.d. having the normal distribution with mean μ and precision τ given (μ, τ). Let (μ, τ) have the usual improper prior. Let $\sigma'^2 = s_n^2/(n-1)$. Prove that the posterior distribution of $V = (n-1)\sigma'^2\tau$ is the χ^2 distribution with $n-1$ degrees of freedom.

5. Suppose that two random variables μ and τ have the joint normal-gamma distribution such that $E(\mu) = -5$. $\text{Var}(\mu) = 1$, $E(\tau) = 1/2$, and $\text{Var}(\tau) = 1/8$. Find the prior hyperparameters μ_0, λ_0, α_0, and β_0 that specify the normal-gamma distribution.

6. Show that two random variables μ and τ cannot have a joint normal-gamma distribution such that $E(\mu) = 0$, $\text{Var}(\mu) = 1$, $E(\tau) = 1/2$, and $\text{Var}(\tau) = 1/4$.

7. Show that two random variables μ and τ cannot have the joint normal-gamma distribution such that $E(\mu) = 0$, $E(\tau) = 1$, and $\text{Var}(\tau) = 4$.

8. Suppose that two random variables μ and τ have the joint normal-gamma distribution with hyperparameters $\mu_0 = 4$, $\lambda_0 = 0.5$, $\alpha_0 = 1$, and $\beta_0 = 8$. Find the values of **(a)** $\Pr(\mu > 0)$ and **(b)** $\Pr(0.736 < \mu < 15.680)$.

9. Using the prior and data in the numerical example on nursing homes in New Mexico in this section, find **(a)** the shortest possible interval such that the posterior probability that μ lies in the interval is 0.90, and **(b)** the shortest possible confidence interval for μ for which the confidence coefficient is 0.90.

10. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown precision τ, and also that the joint prior distribution of μ and τ is the normal-gamma distribution satisfying the following conditions: $E(\mu) = 0$, $E(\tau) = 2$, $E(\tau^2) = 5$, and $\Pr(|\mu| < 1.412) = 0.5$. Determine the prior hyperparameters μ_0, λ_0, α_0, and β_0.

11. Consider again the conditions of Exercise 10. Suppose also that in a random sample of size $n = 10$, it is found that $\overline{x}_n = 1$ and $s_n^2 = 8$. Find the shortest possible interval such that the posterior probability that μ lies in the interval is 0.95.

12. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown precision τ, and also that the joint prior distribution of μ and τ is the normal-gamma distribution satisfying the following conditions: $E(\tau) = 1$, $\text{Var}(\tau) = 1/3$, $\Pr(\mu > 3) = 0.5$, and $\Pr(\mu > 0.12) = 0.9$. Determine the prior hyperparameters μ_0, λ_0, α_0, and β_0.

13. Consider again the conditions of Exercise 12. Suppose also that in a random sample of size $n = 8$, it is found that $\sum_{i=1}^{n} x_i = 16$ and $\sum_{i=1}^{n} x_1^2 = 48$. Find the shortest possible interval such that the posterior probability that μ lies in the interval is 0.99.

14. Continue the analysis in Example 8.6.2 on page 498. Compute an interval (a, b) such that the posterior probability is 0.9 that $a < \mu < b$. Compare this interval with the 90% confidence interval from Example 8.5.4 on page 487.

15. We will draw a sample of size $n = 11$ from the normal distribution with mean μ and precision τ. We will use a natural conjugate prior for the parameters (μ, τ) from the normal-gamma family with hyperparameters $\alpha_0 = 2$, $\beta_0 = 1$, $\mu_0 = 3.5$, and $\lambda_0 = 2$. The sample yields an average of $\overline{x}_n = 7.2$ and $s_n^2 = 20.3$.

a. Find the posterior hyperparameters.

b. Find an interval that contains 95% of the posterior distribution of μ.

16. The study on acid concentration in cheese included a total of 30 lactic acid measurements, the 10 given in Example 8.5.4 on page 487 and the following additional 20:

1.68, 1.9, 1.06, 1.3, 1.52, 1.74, 1.16, 1.49, 1.63, 1.99,
1.15, 1.33, 1.44, 2.01, 1.31, 1.46, 1.72, 1.25, 1.08, 1.25.

a. Using the same prior as in Example 8.6.2 on page 498, compute the posterior distribution of μ and τ based on all 30 observations.

b. Use the posterior distribution found in Example 8.6.2 on page 498 as if it were the prior distribution before observing the 20 observations listed in this problem. Use these 20 new observations to find the posterior distribution of μ and τ and compare the result to the answer to part (a).

17. Consider the analysis performed in Example 8.6.2. This time, use the usual improper prior to compute the posterior distribution of the parameters.

18. Treat the posterior distribution conditional on the first 10 observations found in Exercise 17 as a prior and then observe the 20 additional observations in Exercise 16. Find the posterior distribution of the parameters after observing all of the data and compare it to the distribution found in part (b) of Exercise 16.

19. Consider the situation described in Exercise 7 of Sec. 8.5. Use a prior distribution from the normal-gamma family with values $\alpha_0 = 1$, $\beta_0 = 4$, $\mu_0 = 150$, and $\lambda_0 = 0.5$.

a. Find the posterior distribution of μ and $\tau = 1/\sigma^2$.

b. Find an interval (a, b) such that the posterior probability is 0.90 that $a < \mu < b$.

20. Consider the calorie count data described in Example 7.3.10 on page 400. Now assume that each observation has the normal distribution with unknown mean μ and unknown precision τ given the parameter (μ, τ). Use the normal-gamma conjugate prior distribution with prior hyperparameters $\mu_0 = 0$, $\lambda_0 = 1$, $\alpha_0 = 1$, and $\beta_0 = 60$. The value of s_n^2 is 2102.9.

a. Find the posterior distribution of (μ, τ).

b. Compute $\Pr(\mu > 1|\boldsymbol{x})$.

8.7 Unbiased Estimators

Let δ be an estimator of a function g of a parameter θ. We say that δ is unbiased *if $E_\theta[\delta(\boldsymbol{X})] = g(\theta)$ for all values of θ. This section provides several examples of unbiased estimators.*

Definition of an Unbiased Estimator

Example 8.7.1 Lifetimes of Electronic Components. Consider the company in Example 8.1.3 that wants to estimate the failure rate θ of electronic components. Based on a sample X_1, X_2, X_3 of lifetimes, the M.L.E. of θ is $\hat{\theta} = 3/T$, where $T = X_1 + X_2 + X_3$. The company hopes that $\hat{\theta}$ will be close to θ. The mean of a random variable, such as $\hat{\theta}$, is one measure of where we expect the random variable to be. The mean of $3/T$ is (according to Exercise 21 in Sec. 5.7) $3\theta/2$. If the mean tells us where we expect the estimator to be, we expect this estimator to be 50% larger than θ. ◄

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that involves a parameter (or parameter vector) θ whose value is unknown. Suppose that we wish to estimate a function $g(\theta)$ of the parameter. In a problem of this type, it is desirable to use an estimator $\delta(\boldsymbol{X})$ that, with high probability, will be close to $g(\theta)$. In other words,

it is desirable to use an estimator δ whose distribution changes with the value of θ in such a way that no matter what the true value of θ is, the probability distribution of δ is concentrated around $g(\theta)$.

For example, suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from a normal distribution for which the mean θ is unknown and the variance is 1. In this case, the M.L.E. of θ is the sample mean $\overline{X}_n$. The estimator $\overline{X}_n$ is a reasonably good estimator of θ because its distribution is the normal distribution with mean θ and variance $1/n$. This distribution is concentrated around the unknown value of θ, no matter how large or how small θ is.

These considerations lead to the following definition.

Definition 8.7.1 Unbiased Estimator/Bias. An estimator $\delta(\boldsymbol{X})$ is an *unbiased estimator* of a function $g(\theta)$ of the parameter θ if $E_\theta[\delta(\boldsymbol{X})] = g(\theta)$ for every possible value of θ. An estimator that is not unbiased is called a *biased estimator*. The difference between the expectation of an estimator and $g(\theta)$ is called the *bias* of the estimator. That is, the bias of δ as an estimator of $g(\theta)$ is $E_\theta[\delta(\boldsymbol{X})] - g(\theta)$, and δ is unbiased if and only if the bias is 0 for all θ.

In the case of a sample from a normal distribution with unknown mean θ, $\overline{X}_n$ is an unbiased estimator of θ because $E_\theta(\overline{X}_n) = \theta$ for $-\infty < \theta < \infty$.

Example 8.7.2 Lifetimes of Electronic Components. In Example 8.7.1, the bias of $\hat{\theta} = 3/T$ as an estimator of θ is $3\theta/2 - \theta = \theta/2$. It is easy to see that an unbiased estimator of θ is $\delta(\boldsymbol{X}) = 2/T$. ◄

If an estimator δ of some nonconstant function $g(\theta)$ of the parameter is unbiased, then the distribution of δ must indeed change with the value of θ, since the mean of this distribution is $g(\theta)$. It should be emphasized, however, that this distribution might be either closely concentrated around $g(\theta)$ or widely spread out. For example, an estimator that is equally likely to underestimate $g(\theta)$ by 1,000,000 units or to overestimate $g(\theta)$ by 1,000,000 units would be an unbiased estimator, but it would never yield an estimate close to $g(\theta)$. Therefore, the mere fact that an estimator is unbiased does not necessarily imply that the estimator is good or even reasonable. However, if an unbiased estimator also has a small variance, it follows that the distribution of the estimator will necessarily be concentrated around its mean $g(\theta)$, and there will be high probability that the estimator will be close to $g(\theta)$.

For the reasons just mentioned, the study of unbiased estimators is largely devoted to the search for an unbiased estimator that has a small variance. However, if an estimator δ is unbiased, then its M.S.E. $E_\theta[(\delta - g(\theta))^2]$ is equal to its variance $\text{Var}_\theta(\delta)$. Therefore, the search for an unbiased estimator with a small variance is equivalent to the search for an unbiased estimator with a small M.S.E. The following result is a simple corollary to Exercise 4 in Sec. 4.3.

Corollary 8.7.1 Let δ be an estimator with finite variance. Then the M.S.E. of δ as an estimator of $g(\theta)$ equals its variance plus the square of its bias.

Example 8.7.3 Lifetimes of Electronic Components. We can compare the two estimators $\hat{\theta}$ and $\delta(\boldsymbol{X})$ in Example 8.7.2 using M.S.E. According to Exercise 21 in Sec. 5.7, the variance of $1/T$ is $\theta^2/4$. So, the M.S.E. of $\delta(\boldsymbol{X})$ is θ^2. For $\hat{\theta}$, the variance is $9\theta^2/4$ and the square of the bias is $\theta^2/4$, so the M.S.E. is $5\theta^2/2$, which is 2.5 times as large as the M.S.E. of $\delta(\boldsymbol{X})$. If M.S.E. were the sole concern, the estimator $\delta^*(\boldsymbol{X}) = 1/T$ has variance

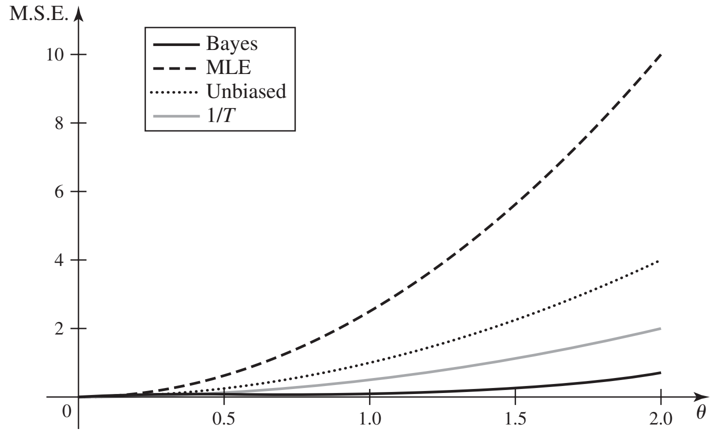


Figure 8.8 M.S.E. for each of the four estimators in Example 8.7.3.

and squared bias both equal to $\theta^2/4$, so the M.S.E. is $\theta^2/2$, half the M.S.E. of the unbiased estimator. Figure 8.8 plots the M.S.E. for each of these estimators together with the M.S.E. of the Bayes estimator $4/(2+T)$ found in Example 8.1.3. Calculation of the M.S.E. of the Bayes estimator required simulation. Eventually (above $\theta = 3.1$), the M.S.E. of the Bayes estimator crosses above the M.S.E. of $1/T$, but it stays below the other two for all θ. ◄

Example 8.7.4 Unbiased Estimation of the Mean. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter (or parameter vector) θ. Assume that the mean and variance of the distribution are finite. Define $g(\theta) = E_\theta(X_1)$. The sample mean $\overline{X}_n$ is obviously an unbiased estimator of $g(\theta)$. Its M.S.E. is $\text{Var}_\theta(X_1)/n$. In Example 8.7.1, $g(\theta) = 1/\theta$ and $\overline{X}_n = 1/\hat{\theta}$ is an unbiased estimator the mean. ◄

Unbiased Estimation of the Variance

Theorem 8.7.1 Sampling from a General Distribution. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter (or parameter vector) θ. Assume that the variance of the distribution is finite. Define $g(\theta) = \text{Var}_\theta(X_1)$. The following statistic is an unbiased estimator of the variance $g(\theta)$:

$$\hat{\sigma}_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

Proof Let $\mu = E_\theta(X_1)$, and let σ^2 stand for $g(\theta) = \text{Var}_\theta(X_1)$. Since the sample mean is an unbiased estimator of μ, it is more or less natural to consider first the sample variance $\hat{\sigma}_0^2 = (1/n)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ and to attempt to determine if it is an unbiased estimator of the variance σ^2. We shall use the identity

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + n(\overline{X}_n - \mu)^2.$$

Then it follows that

$$E(\hat{\sigma}_0^2) = E\left[\frac{1}{n}\sum\nolimits_{i=1}^{n}(X_i - \overline{X}_n)^2\right]$$
$$= E\left[\frac{1}{n}\sum\nolimits_{i=1}^{n}(X_i - \mu)^2\right] - E[(\overline{X}_n - \mu)^2]. \tag{8.7.1}$$

Since each observation X_i has mean μ and variance σ^2, then $E[(X_i - \mu)^2] = \sigma^2$ for $i = 1, \ldots, n$. Therefore,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}E[(X_i - \mu)^2] = \frac{1}{n}n\sigma^2 = \sigma^2. \tag{8.7.2}$$

Furthermore, the sample mean $\overline{X}_n$ has mean μ and variance σ^2/n. Therefore,

$$E[(\overline{X}_n - \mu)^2] = \text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}. \tag{8.7.3}$$

It now follows from Eqs. (8.7.1), (8.7.2), and (8.7.3) that

$$E(\hat{\sigma}_0^2) = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2. \tag{8.7.4}$$

It can be seen from Eq. (8.7.4) that the sample variance $\hat{\sigma}_0^2$ is not an unbiased estimator of σ^2, because its expectation is $[(n-1)/n]\sigma^2$, rather than σ^2. However, if $\hat{\sigma}_0^2$ is multiplied by the factor $n/(n-1)$ to obtain the statistic $\hat{\sigma}_1^2$, then the expectation of $\hat{\sigma}_1^2$ will indeed be σ^2. Therefore, $\hat{\sigma}_1^2$ is an unbiased estimator of σ^2. ■

In light of Theorem 8.7.1, many textbooks define the sample variance as $\hat{\sigma}_1^2$, rather than as $\hat{\sigma}_0^2$.

Note: Special Case of Normal Random Sample. The estimator $\hat{\sigma}_0^2$ is the same as the maximum likelihood estimator $\widehat{\sigma^2}$ of σ^2 when $X_1, \ldots, X_n$ have the normal distribution with mean μ and variance σ^2. Also, $\hat{\sigma}_1^2$ is the same as the random variable σ'^2 that appears in confidence intervals for μ. We have chosen to use different names for these estimators in this section because we are discussing general distributions for which σ^2 might be some function $g(\theta)$ whose M.L.E. is completely different from $\hat{\sigma}_0^2$. (See Exercise 1 for one such example.)

Sampling from a Specific Family of Distributions When it can be assumed that $X_1, \ldots, X_n$ form a random sample from a specific family of distributions, such as the family of Poisson distributions, it will generally be desirable to consider not only $\hat{\sigma}_1^2$ but also other unbiased estimators of the variance.

Example 8.7.5 Sample from a Poisson Distribution. Suppose that we observe a random sample from the Poisson distribution for which the mean θ is unknown. We have already seen that $\overline{X}_n$ will be an unbiased estimator of the mean θ. Moreover, since the variance of a Poisson distribution is also equal to θ, it follows that $\overline{X}_n$ is also an unbiased estimator of the variance. In this example, therefore, both $\overline{X}_n$ and $\hat{\sigma}_1^2$ are unbiased estimators of the unknown variance θ. Furthermore, any combination of $\overline{X}_n$ and $\hat{\sigma}_1^2$ having the form $\alpha\overline{X}_n + (1-\alpha)\hat{\sigma}_1^2$, where α is a given constant $(-\infty < \alpha < \infty)$, will also be an unbiased estimator of θ because its expectation will be

$$E[\alpha\overline{X}_n + (1-\alpha)\hat{\sigma}_1^2] = \alpha E(\overline{X}_n) + (1-\alpha)E(\hat{\sigma}_1^2) = \alpha\theta + (1-\alpha)\theta = \theta. \tag{8.7.5}$$

Other unbiased estimators of θ can also be constructed. ◂

If an unbiased estimator is to be used, the problem is to determine which one of the possible unbiased estimators has the smallest variance or, equivalently, has the smallest M.S.E. We shall not derive the solution to this problem right now. However, it will be shown in Sec. 8.8 that in Example 8.7.5, for every possible value of θ, the estimator $\overline{X}_n$ has the smallest variance among all unbiased estimators of θ. This result is not surprising. We know from Example 7.7.2 that $\overline{X}_n$ is a sufficient statistic for θ, and it was argued in Sec. 7.9 that we can restrict our attention to estimators that are functions of the sufficient statistic alone. (See also Exercise 13 at the end of this section.)

Example 8.7.6 Sampling from a Normal Distribution. Assume that $\boldsymbol{X} = (X_1, \dots, X_n)$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. We shall consider the problem of estimating σ^2. We know from Theorem 8.7.1 that the estimator $\hat{\sigma}_1^2$ is an unbiased estimator of σ^2. Moreover, we know from Example 7.5.6 that the sample variance $\hat{\sigma}_0^2$ is the M.L.E. of σ^2. We want to determine whether the M.S.E. $E[(\hat{\sigma}_i^2 - \sigma^2)^2]$ is smaller for the estimator $\hat{\sigma}_0^2$ or for the estimator $\hat{\sigma}_1^2$, and also whether or not there is some other estimator of σ^2 that has a smaller M.S.E. than both $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$.

Both the estimator $\hat{\sigma}_0^2$ and the estimator $\hat{\sigma}_1^2$ have the following form:

$$T_c = c\sum_{i=1}^{n}(X_i - \overline{X}_n)^2, \tag{8.7.6}$$

where $c = 1/n$ for $\hat{\sigma}_0^2$ and $c = 1/(n-1)$ for $\hat{\sigma}_1^2$. We shall now determine the M.S.E. for an arbitrary estimator having the form in Eq. (8.7.6) and shall then determine the value of c for which this M.S.E. is minimum. We shall demonstrate the striking property that the same value of c minimizes the M.S.E. for all possible values of the parameters μ and σ^2. Therefore, among all estimators having the form in Eq. (8.7.6), there is a single one that has the smallest M.S.E. for all possible values of μ and σ^2.

It was shown in Sec. 8.3 that when $X_1, \dots, X_n$ form a random sample from a normal distribution, the random variable $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2/\sigma^2$ has the χ^2 distribution with $n-1$ degrees of freedom. By Theorem 8.2.1, the mean of this variable is $n-1$, and the variance is $2(n-1)$. Therefore, if T_c is defined by Eq. (8.7.6), then

$$E(T_c) = (n-1)c\sigma^2 \quad \text{and} \quad \text{Var}(T_c) = 2(n-1)c^2\sigma^4. \tag{8.7.7}$$

Thus, by Corollary 8.7.1, the M.S.E. of T_c can be found as follows:

$$\begin{aligned} E[(T_c - \sigma^2)^2] &= [E(T_c) - \sigma^2]^2 + \text{Var}(T^2) \\ &= [(n-1)c - 1]^2\sigma^4 + 2(n-1)c^2\sigma^4 \\ &= [(n^2-1)c^2 - 2(n-1)c + 1]\sigma^4. \end{aligned} \tag{8.7.8}$$

The coefficient of σ^4 in Eq. (8.7.8) is simply a quadratic function of c. Hence, no matter what σ^2 equals, the minimizing value of c is found by elementary differentiation to be $c = 1/(n+1)$.

In summary, we have established the following fact: Among all estimators of σ^2 having the form in Eq. (8.7.6), the estimator that has the smallest M.S.E. for all possible values of μ and σ^2 is $T_{1/(n+1)} = [1/(n+1)]\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. In particular, $T_{1/(n+1)}$ has a smaller M.S.E. than both the M.L.E. $\hat{\sigma}_0^2$ and the unbiased estimator $\hat{\sigma}_1^2$. Therefore, the estimators $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$, as well as all other estimators having the form in Eq. (8.7.6) with $c \neq 1/(n+1)$, are inadmissible. Furthermore, it was shown

by C. Stein in 1964 that even the estimator $T_{1/(n+1)}$ is dominated by other estimators and that $T_{1/(n+1)}$ itself is therefore inadmissible.

The estimators $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ are compared in Exercise 6 at the end of this section. Of course, when the sample size n is large, it makes little difference whether n, $n-1$, or $n+1$ is used as the divisor in the estimate of σ^2; all three estimators $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$, and $T_{1/(n+1)}$ will be approximately equal. ◀

Limitations of Unbiased Estimation

The concept of unbiased estimation has played an important part in the historical development of statistics, and the feeling that an unbiased estimator should be preferred to a biased estimator is prevalent in current statistical practice. Indeed, what scientist wishes to be biased or to be accused of being biased? The very terminology of the theory of unbiased estimation seems to make the use of unbiased estimators highly desirable.

However, as explained in this section, the quality of an unbiased estimator must be evaluated in terms of its variance or its M.S.E. Examples 8.7.3 and 8.7.6 illustrate the following fact: In many problems, there exist biased estimators that have smaller M.S.E. than every unbiased estimator for every possible value of the parameter. Furthermore, it can be shown that a Bayes estimator, which makes use of all relevant prior information about the parameter and which minimizes the overall M.S.E., is unbiased only in trivial problems in which the parameter can be estimated perfectly.

Some other limitations of the theory of unbiased estimation will now be described.

Nonexistence of an Unbiased Estimator In many problems, there does not exist any unbiased estimator of the function of the parameter that must be estimated. For example, suppose that $X_1, \ldots, X_n$ form n Bernoulli trials for which the parameter p is unknown ($0 \le p \le 1$). Then the sample mean $\overline{X}_n$ will be an unbiased estimator of p, but it can be shown that there will be no unbiased estimator of $p^{1/2}$. (See Exercise 7.) Furthermore, if it is known in this example that p must lie in the interval $\frac{1}{3} \le p \le \frac{2}{3}$, then there is no unbiased estimator of p whose possible values are confined to that same interval.

Inappropriate Unbiased Estimators Consider an infinite sequence of Bernoulli trials for which the parameter p is unknown ($0 < p < 1$), and let X denote the number of failures that occur before the first success is obtained. Then X has the geometric distribution with parameter p whose p.f. is given by Eq. (5.5.3). If it is desired to estimate the value of p from the observation X, then it can be shown (see Exercise 8) that the *only* unbiased estimator of p yields the estimate 1 if $X = 0$ and yields the estimate 0 if $X > 0$. This estimator seems inappropriate. For example, if the first success is obtained on the second trial, that is, if $X = 1$, then it is silly to estimate that the probability of success p is 0. Similarly, if $X = 0$ (the first trial is success), it seems silly to estimate p to be as large as 1.

As another example of an inappropriate unbiased estimator, suppose that the random variable X has the Poisson distribution with unknown mean λ ($\lambda > 0$), and suppose also that it is desired to estimate the value of $e^{-2\lambda}$. It can be shown (see Exercise 9) that the *only* unbiased estimator of $e^{-2\lambda}$ yields the estimate 1 if X is an even integer and the estimate -1 if X is an odd integer. This estimator is inappropriate for two reasons. First, it yields the estimate 1 or -1 for a parameter $e^{-2\lambda}$, which must

lie between 0 and 1. Second, the value of the estimate depends only on whether X is odd or even, rather than on whether X is large or small.

Ignoring Information One more criticism of the concept of unbiased estimation is that the principle of always using an unbiased estimator for a parameter θ (when such exists) sometimes ignores valuable information that is available. As an example, suppose that the average voltage θ in a certain electric circuit is unknown; this voltage is to be measured by a voltmeter for which the reading X has the normal distribution with mean θ and known variance σ^2. Suppose also that the observed reading on the voltmeter is 2.5 volts. Since X is an unbiased estimator of θ in this example, a scientist who wished to use an unbiased estimator would estimate the value of θ to be 2.5 volts.

However, suppose also that after the scientist reported the value 2.5 as his estimate of θ, he discovered that the voltmeter actually truncates all readings at 3 volts, just as in Example 3.2.7 on page 106. That is, the reading of the voltmeter is accurate for any voltage less than 3 volts, but a voltage greater than 3 volts would be reported as 3 volts. Since the actual reading was 2.5 volts, this reading was unaffected by the truncation. Nevertheless, the observed reading would no longer be an unbiased estimator of θ because the distribution of the truncated reading X is not a normal distribution with mean θ. Therefore, if the scientist still wished to use an unbiased estimator, he would have to change his estimate of θ from 2.5 volts to a different value.

Ignoring the fact that the observed reading was accurate seems unacceptable. Since the actual observed reading was only 2.5 volts, it is the same as what would have been observed if there had been no truncation. Since the observed reading is untruncated, it would seem that the fact that there might have been a truncated reading is irrelevant to the estimation of θ. However, since this possibility does change the sample space of X and its probability distribution, it will also change the form of the unbiased estimator of θ.

Summary

An estimator $\delta(\boldsymbol{X})$ of $g(\theta)$ is unbiased if $E_\theta[\delta(\boldsymbol{X})] = g(\theta)$ for all possible values of θ. The bias of an estimator of $g(\theta)$ is $E_\theta[\delta(\boldsymbol{X})] - g(\theta)$. The M.S.E. of an estimator equals its variance plus the square of its bias. The M.S.E. of an unbiased estimator equals its variance.

Exercises

1. Let $X_1, \ldots, X_n$ be a random sample from the Poisson distribution with mean θ.

a. Express the $\text{Var}_\theta(X_i)$ as a function $\sigma^2 = g(\theta)$.

b. Find the M.L.E. of $g(\theta)$ and show that it is unbiased.

2. Suppose that X is a random variable whose distribution is completely unknown, but it is known that all the moments $E(X^k)$, for $k = 1, 2, \ldots$, are finite. Suppose also that $X_1, \ldots, X_n$ form a random sample from this distribution. Show that for $k = 1, 2, \ldots$, the kth sample moment $(1/n)\sum_{i=1}^n X_i^k$ is an unbiased estimator of $E(X^k)$.

3. For the conditions of Exercise 2, find an unbiased estimator of $[E(X)]^2$. *Hint:* $[E(X)]^2 = E(X^2) - \text{Var}(X)$.

4. Suppose that a random variable X has the geometric distribution with unknown parameter p. (See Sec. 5.5.) Find a statistic $\delta(X)$ that will be an unbiased estimator of $1/p$.

5. Suppose that a random variable X has the Poisson distribution with unknown mean λ ($\lambda > 0$). Find a statistic $\delta(X)$ that will be an unbiased estimator of e^{λ}. *Hint:* If $E[\delta(X)] = e^{\lambda}$, then

$$\sum_{x=0}^{\infty} \frac{\delta(x)e^{-\lambda}\lambda^x}{x!} = e^{\lambda}.$$

Multiply both sides of this equation by e^{λ}, expand the right side in a power series in λ, and then equate the coefficients of λ^x on both sides of the equation for $x = 0, 1, 2, \dots$.

6. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Let $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ be the two estimators of σ^2, which are defined as follows:

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \text{ and } \hat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i-1}^{n}(X_i - \overline{X}_n)^2.$$

Show that the M.S.E. of $\hat{\sigma}_0^2$ is smaller than the M.S.E. of $\hat{\sigma}_1^2$ for all possible values of μ and σ^2.

7. Suppose that $X_1, \dots, X_n$ form n Bernoulli trials for which the parameter p is unknown ($0 \le p \le 1$). Show that the expectation of every function $\delta(X_1, \dots, X_n)$ is a polynomial in p whose degree does not exceed n.

8. Suppose that a random variable X has the geometric distribution with unknown parameter p ($0 < p < 1$). Show that the only unbiased estimator of p is the estimator $\delta(X)$ such that $\delta(0) = 1$ and $\delta(X) = 0$ for $X > 0$.

9. Suppose that a random variable X has the Poisson distribution with unknown mean λ ($\lambda > 0$). Show that the only unbiased estimator of $e^{-2\lambda}$ is the estimator $\delta(X)$ such that $\delta(X) = 1$ if X is an even integer and $\delta(X) = -1$ if X is an odd integer.

10. Consider an infinite sequence of Bernoulli trials for which the parameter p is unknown ($0 < p < 1$), and suppose that sampling is continued until exactly k successes have been obtained, where k is a fixed integer ($k \ge 2$). Let N denote the total number of trials that are needed to obtain the k successes. Show that the estimator $(k-1)/(N-1)$ is an unbiased estimator of p.

11. Suppose that a certain drug is to be administered to two different types of animals A and B. It is known that the mean response of animals of type A is the same as the mean response of animals of type B, but the common value θ of this mean is unknown and must be estimated. It is also known that the variance of the response of animals of type A is four times as large as the variance of the response of animals of type B. Let $X_1, \dots, X_m$ denote the responses of a random sample of m animals of type A, and let $Y_1, \dots, Y_n$ denote the responses of an independent random sample of n animals of type B. Finally, consider the estimator $\hat{\theta} = \alpha\overline{X}_m + (1-\alpha)\overline{Y}_n$.

a. For what values of α, m, and n is $\hat{\theta}$ an unbiased estimator of θ?

b. For fixed values of m and n, what value of α yields an unbiased estimator with minimum variance?

12. Suppose that a certain population of individuals is composed of k different strata ($k \ge 2$), and that for $i = 1, \dots, k$, the proportion of individuals in the total population who belong to stratum i is p_i, where $p_i > 0$ and $\sum_{i=1}^{k} p_i = 1$. We are interested in estimating the mean value μ of a certain characteristic among the total population. Among the individuals in stratum i, this characteristic has mean μ_i and variance σ_i^2, where the value of μ_i is unknown and the value of σ_i^2 is known. Suppose that a *stratified sample* is taken from the population as follows: From each stratum i, a random sample of n_i individuals is taken, and the characteristic is measured for each of these individuals. The samples from the k strata are taken independently of each other. Let $\overline{X}_i$ denote the average of the n_i measurements in the sample from stratum i.

a. Show that $\mu = \sum_{i=1}^{k} p_i\mu_i$, and show also that $\hat{\mu} = \sum_{i=1}^{k} p_i\overline{X}_i$ is an unbiased estimator of μ.

b. Let $n = \sum_{i=1}^{k} n_i$ denote the total number of observations in the k samples. For a fixed value of n, find the values of $n_1, \dots, n_k$ for which the variance of $\hat{\mu}$ will be a minimum.

13. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where the value of the parameter θ is unknown. Let $\boldsymbol{X} = (X_1, \dots, X_n)$, and let T be a statistic. Assume that $\delta(\boldsymbol{X})$ is an unbiased estimator of θ such that $E_\theta[\delta(\boldsymbol{X})|T]$ does not depend on θ. (If T is a sufficient statistic, as defined in Sec. 7.7, then this will be true for every estimator δ. The condition also holds in other examples.) Let $\delta_0(T)$ denote the conditional mean of $\delta(\boldsymbol{X})$ given T.

a. Show that $\delta_0(T)$ is also an unbiased estimator of θ.

b. Show that $\text{Var}_\theta(\delta_0) \le \text{Var}_\theta(\delta)$ for every possible value of θ. *Hint:* Use the result of Exercise 11 in Sec. 4.7.

14. Suppose that $X_1, \dots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where the value of the parameter θ is unknown; and let $Y_n = \max(X_1, \dots, X_n)$. Show that $[(n+1)/n]Y_n$ is an unbiased estimator of θ.

15. Suppose that a random variable X can take only the five values $x = 1, 2, 3, 4, 5$ with the following probabilities:

$$f(1|\theta) = \theta^3, \quad f(2|\theta) = \theta^2(1-\theta),$$
$$f(3|\theta) = 2\theta(1-\theta), \; f(4|\theta) = \theta(1-\theta)^2,$$
$$f(5|\theta) = (1-\theta)^3.$$

Here, the value of the parameter θ is unknown ($0 \le \theta \le 1$).

a. Verify that the sum of the five given probabilities is 1 for every value of θ.

b. Consider an estimator $\delta_c(X)$ that has the following form:

$$\delta_c(1) = 1,\ \delta_c(2) = 2 - 2c,\ \delta_c(3) = c,$$
$$\delta_c(4) = 1 - 2c,\ \delta_c(5) = 0.$$

Show that for each constant c, $\delta_c(X)$ is an unbiased estimator of θ.

c. Let θ_0 be a number such that $0 < \theta_0 < 1$. Determine a constant c_0 such that when $\theta = \theta_0$, the variance of $\delta_{c_0}(X)$ is smaller than the variance of $\delta_c(X)$ for every other value of c.

16. Reconsider the conditions of Exercise 3. Suppose that $n = 2$, and we observe $X_1 = 2$ and $X_2 = -1$. Compute the value of the unbiased estimator of $[E(X)]^2$ found in Exercise 3. Describe a flaw that you have discovered in the estimator.

⋆ 8.8 Fisher Information

This section introduces a method for measuring the amount of information that a sample of data contains about an unknown parameter. This measure has the intuitive properties that more data provide more information, and more precise data provide more information. The information measure can be used to find bounds on the variances of estimators, and it can be used to approximate the variances of estimators obtained from large samples.

Definition and Properties of Fisher Information

Example 8.8.1 Studying Customer Arrivals. A store owner is interested in learning about customer arrivals. She models arrivals during the day as a Poisson process (see Definition 5.4.2) with unknown rate θ. She thinks of two different possible sampling plans to obtain information about customer arrivals. One plan is to choose a fixed number, n, of customers and to see how long, X, it takes until n customers arrive. The other plan is to observe for a fixed length of time, t, and count how many customers, Y, arrive during time t. That is, the store owner can either observe a Poisson random variable, Y, with mean $t\theta$ or observe a gamma random variable, X, with parameters n and θ. Is there any way to address the question of which sampling plan is likely to be more informative? ◀

The Fisher information is one property of a distribution that can be used to measure how much information one is likely to obtain from a random variable or a random sample.

The Fisher Information in a Single Random Variable In this section, we shall introduce a concept, called the Fisher information, that enters various aspects of the theory of statistical inference, and we shall describe a few uses of this concept.

Consider a random variable X for which the p.f. or the p.d.f. is $f(x|\theta)$. It is assumed that $f(x|\theta)$ involves a parameter θ whose value is unknown but must lie in a given open interval Ω of the real line. Furthermore, it is assumed that X takes values in a specified sample space S, and $f(x|\theta) > 0$ for each value of $x \in S$ and each value of $\theta \in \Omega$. This assumption eliminates from consideration the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown, because, for that distribution, $f(x|\theta) > 0$ only when $x < \theta$ and $f(x|\theta) = 0$ when $x > \theta$. The assumption does not eliminate any distribution where the set of values of x for which $f(x|\theta) > 0$ is a fixed set that does not depend on θ.

Next, we define $\lambda(x|\theta)$ as follows:

$$\lambda(x|\theta) = \log f(x|\theta).$$

It is assumed that for each value of $x \in S$, the p.f. or p.d.f. $f(x|\theta)$ is a twice differentiable function of θ, and we let

$$\lambda'(x|\theta) = \frac{\partial}{\partial\theta}\lambda(x|\theta) \quad \text{and} \quad \lambda''(x|\theta) = \frac{\partial^2}{\partial\theta^2}\lambda(x|\theta).$$

Definition 8.8.1 Fisher Information in a Random Variable. Let X be a random variable whose distribution depends on a parameter θ that takes values in an open interval Ω of the real line. Let the p.f. or p.d.f. of X be $f(x|\theta)$. Assume that the set of x such that $f(x|\theta) > 0$ is the same for all θ and that $\lambda(x|\theta) = \log f(x|\theta)$ is twice differentiable as a function of θ. The *Fisher information* $I(\theta)$ in the random variable X is defined as

$$I(\theta) = E_\theta\{[\lambda'(X|\theta)]^2\}. \tag{8.8.1}$$

Thus, if $f(x|\theta)$ is a p.d.f., then

$$I(\theta) = \int_S [\lambda'(x|\theta)]^2 f(x|\theta)\, dx. \tag{8.8.2}$$

If $f(x|\theta)$ is a p.f., the integral in Eq. (8.8.2) is replaced by a sum over the points in S. In the discussion that follows, we shall assume for convenience that $f(x|\theta)$ is a p.d.f. However, all the results hold also when $f(x|\theta)$ is a p.f.

An alternative method for calculating the Fisher information sometimes proves more useful.

Theorem 8.8.1 Assume the conditions of Definition 8.8.1. Also, assume that two derivatives of $\int_S f(x|\theta)dx$ with respect to θ can be calculated by reversing the order of integration and differentiation. Then the Fisher information also equals

$$I(\theta) = -E_\theta[\lambda''(X|\theta)]. \tag{8.8.3}$$

Another expression for the Fisher information is

$$I(\theta) = \text{Var}_\theta[\lambda'(X|\theta)]. \tag{8.8.4}$$

Proof We know that $\int_S f(x|\theta)\, dx = 1$ for every value of $\theta \in \Omega$. Therefore, if the integral on the left side of this equation is differentiated with respect to θ, the result will be 0. We have assumed that we can reverse the order in which we perform the integration with respect to x, and the differentiation with respect to θ, and will still obtain the value 0. In other words, we shall assume that we can take the derivative inside the integral sign and obtain

$$\int_S f'(x|\theta)\, dx = 0 \quad \text{for } \theta \in \Omega. \tag{8.8.5}$$

Furthermore, we have assumed that we can take a second derivative with respect to θ "inside the integral sign" and obtain

$$\int_S f''(x|\theta)\, dx = 0 \quad \text{for } \theta \in \Omega. \tag{8.8.6}$$

Since $\lambda'(x|\theta) = f'(x|\theta)/f(x|\theta)$, then

$$E_\theta[\lambda'(X|\theta)] = \int_S \lambda'(x|\theta) f(x|\theta)\, dx = \int_S f'(x|\theta)\, dx.$$

Hence, it follows from Eq. (8.8.5) that

$$E_\theta[\lambda'(X|\theta)] = 0. \tag{8.8.7}$$

Since the mean of $\lambda'(X|\theta)$ is 0, it follows from Eq. (8.8.1) that Eq. (8.8.4) holds.

Next, note that

$$\lambda''(x|\theta) = \frac{f(x|\theta)f''(x|\theta) - [f'(x|\theta)]^2}{[f(x|\theta)]^2}$$
$$= \frac{f''(x|\theta)}{f(x|\theta)} - [\lambda'(x|\theta)]^2.$$

Therefore,

$$E_\theta[\lambda''(X|\theta)] = \int_S f''(x|\theta)\,dx - I(\theta). \tag{8.8.8}$$

It follows from Eqs. (8.8.8) and (8.8.6) that Eq. (8.8.3) holds. ■

In many problems, it is easier to determine the value of $I(\theta)$ from Eq. (8.8.3) than from Eqs. (8.8.1) or (8.8.4).

Example 8.8.2

The Bernoulli Distributions. Suppose that X has the Bernoulli distribution with parameter p. We shall determine the Fisher information $I(p)$ in X.

In this example, the possible values of X are the two values 0 and 1. For $x = 0$ or 1,

$$\lambda(x|p) = \log f(x|p) = x \log p + (1 - x)\log(1 - p).$$

Hence,

$$\lambda'(x|p) = \frac{x}{p} - \frac{1 - x}{1 - p}$$

and

$$\lambda''(x|p) = -\left[\frac{x}{p^2} + \frac{1 - x}{(1 - p)^2}\right].$$

Since $E(X) = p$, the Fisher information is

$$I(p) = -E[\lambda''(X|p)] = \frac{1}{p} + \frac{1}{1 - p} = \frac{1}{p(1 - p)}.$$

Recall from Eq. (4.3.3) that $\text{Var}(X) = p(1 - p)$, so the more precise (smaller variance) X is the more information it provides.

In this example, it can be readily verified that the assumptions made in the proof of Theorem 8.8.1 are satisfied. Indeed, because X can take only the two values 0 and 1, the integrals in Eqs. (8.8.5) and (8.8.6) reduce to summations over the two values $x = 0$ and $x = 1$. Since it is always possible to take a derivative "inside a finite summation" and to differentiate the sum term by term, Eqs. (8.8.5) and (8.8.6) must be satisfied. ◀

Example 8.8.3

The Normal Distributions. Suppose that X has the normal distribution with unknown mean μ and known variance σ^2. We shall determine the Fisher information $I(\mu)$ in X.

For $-\infty < x < \infty$,

$$\lambda(x|\mu) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x - \mu)^2}{2\sigma^2}.$$

Hence,

$$\lambda'(x|\mu) = \frac{x-\mu}{\sigma^2} \quad \text{and} \quad \lambda''(x|\mu) = -\frac{1}{\sigma^2}.$$

It now follows from Eq. (8.8.3) that the Fisher information is

$$I(\mu) = \frac{1}{\sigma^2}.$$

Since $\text{Var}(X) = \sigma^2$, we see again that the more precise (smaller variance) X is, the more information it provides.

In this example, it can be verified directly (see Exercise 1 at the end of this section) that Eqs. (8.8.5) and (8.8.6) are satisfied. ◀

It should be emphasized that the concept of Fisher information cannot be applied to a distribution, such as the uniform distribution on the interval $[0, \theta]$, for which the necessary assumptions are not satisfied.

The Fisher Information in a Random Sample When we have a random sample from a distribution, the Fisher information is defined in an analogous manner. Indeed, Definition 8.8.2 subsumes Definition 8.8.1 as the special case in which $n = 1$.

Definition 8.8.2 Fisher Information in a Random Sample. Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from a distribution for which the p.f. or p.d.f. is $f(x|\theta)$, where the value of the parameter θ must lie in an open interval Ω of the real line. Let $f_n(\boldsymbol{x}|\theta)$ denote the joint p.f. or joint p.d.f. of $\boldsymbol{X}$. Define

$$\lambda_n(\boldsymbol{x}|\theta) = \log f_n(x|\theta). \tag{8.8.9}$$

Assume that the set of $\boldsymbol{x}$ such that $f_n(\boldsymbol{x}|\theta) > 0$ is the same for all θ and that $\log f_n(\boldsymbol{x}|\theta)$ is twice differentiable with respect to θ. The *Fisher information* $I_n(\theta)$ in the random sample $\boldsymbol{X}$ is defined as

$$I_n(\theta) = E_\theta\{[\lambda_n'(\boldsymbol{X}|\theta)]^2\}.$$

For continuous distributions, the Fisher information $I_n(\theta)$ in the entire sample is given by the following n-dimensional integral:

$$I_n(\theta) = \int_S \cdots \int_S [\lambda_n'(\boldsymbol{x}|\theta)]^2 f_n(\boldsymbol{x}|\theta)\, dx_1 \ldots dx_n.$$

For discrete distributions, replace the n-dimensional integral by an n-fold summation.

Furthermore, if we again assume that derivatives can be passed under the integrals, then we may express $I_n(\theta)$ in either of the following two ways:

$$I_n(\theta) = \text{Var}_\theta[\lambda_n'(\boldsymbol{X}|\theta)] \tag{8.8.10}$$

or

$$I_n(\theta) = -E_\theta[\lambda_n''(\boldsymbol{X}|\theta)]. \tag{8.8.11}$$

We shall now show that there is a simple relation between the Fisher information $I_n(\theta)$ in the entire sample and the Fisher information $I(\theta)$ in a single observation X_i.

Theorem 8.8.2 Under the conditions of Definitions 8.8.1 and 8.8.2,

$$I_n(\theta) = nI(\theta). \tag{8.8.12}$$

In words, the Fisher information in a random sample of n observations is simply n times the Fisher information in a single observation.

Proof Since $f_n(\boldsymbol{x}|\theta) = f(x_1|\theta) \ldots f(x_n|\theta)$, it follows that

$$\lambda_n(\boldsymbol{x}|\theta) = \sum_{i=1}^{n} \lambda(x_i|\theta).$$

Hence,

$$\lambda_n''(\boldsymbol{x}|\theta) = \sum_{i=1}^{n} \lambda''(x_i|\theta). \tag{8.8.13}$$

Since each observation X_i has the p.d.f. $f(x|\theta)$, the Fisher information in each X_i is $I(\theta)$. It follows from Eqs. (8.8.3) and (8.8.11) that by taking expectations on both sides of Eq. (8.8.13), we obtain Eq. (8.8.12). ■

Example 8.8.4 Studying Customer Arrivals. Return to the store owner in Example 8.8.1 who is trying to choose between sampling a Poisson random variable, Y, with mean $t\theta$ or sampling a gamma random variable, X, with parameters n and θ. The reader can compute the Fisher information in each random variable in Exercises 3 and 19 in this section. We shall label them $I_Y(\theta)$ and $I_X(\theta)$. They are

$$I_X(\theta) = \frac{n}{\theta^2} \quad \text{and} \quad I_Y(\theta) = \frac{t}{\theta}.$$

Which is larger will clearly depend on the particular values of n, t, and θ. Both n and t can be chosen by the store owner, but θ is unknown. In order for $I_X(\theta) = I_Y(\theta)$, it is necessary and sufficient that $n = t\theta$. This relation actually makes intuitive sense. For example, if the store owner chooses to observe Y, then the total number N of customers observed will be random and $N = Y$. The mean of N is then $E(Y) = t\theta$. Similarly, if the store owner chooses to observe X, then the length of time T that it takes to observe n customers will be random. In fact, $T = X$, and the mean of $T\theta$ is n. So long as the manufacturer is comparing sampling plans that are expected to observe the same numbers of customers or observe for the same length of time, the two sampling plans should provide the same amount of information. ◄

The Information Inequality

Example 8.8.5 Studying Customer Arrivals. Another way that the store owner in Example 8.8.4 could choose between the two sampling plans is to compare the estimators that she will use to make inferential statements about customer arrivals. For example, she may want to estimate θ, the rate of customer arrivals. Alternatively, she may want to estimate $1/\theta$, the mean time between customer arrivals. Each sampling plan lends itself to estimation of both parameters. Indeed, there are unbiased estimators of both parameters available from at least one of these sampling plans. ◄

As one application of the results that have been derived concerning Fisher information, we shall show how the Fisher information can be used to determine a lower bound for the variance of an arbitrary estimator of the parameter θ in a given problem. The following result was independently developed by H. Cramér and C. R. Rao during the 1940s.

Theorem 8.8.3 Cramér-Rao (Information) Inequality. Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from a distribution for which the p.d.f. is $f(x|\theta)$. Suppose also that all the

assumptions which have been made about $f(x|\theta)$ thus far in this section continue to hold. Let $T = r(\boldsymbol{X})$ be a statistic with finite variance. Let $m(\theta) = E_\theta(T)$. Assume that $m(\theta)$ is a differentiable function of θ. Then

$$\mathrm{Var}_\theta(T) \geq \frac{[m'(\theta)]^2}{nI(\theta)}. \tag{8.8.14}$$

There will be equality in (8.8.14) if and only if there exist functions $u(\theta)$ and $v(\theta)$ that may depend on θ but do not depend on $\boldsymbol{X}$ and that satisfy the relation

$$T = u(\theta)\lambda_n'(\boldsymbol{X}|\theta) + v(\theta). \tag{8.8.15}$$

Proof The inequality derives from applying Theorem 4.6.3 to the covariance between T and the random variable $\lambda_n'(\boldsymbol{X}|\theta)$ defined in Eq. (8.8.9). Since $\lambda_n'(\boldsymbol{x}|\theta) = f_n'(\boldsymbol{x}|\theta)/f_n(\boldsymbol{x}|\theta)$, it follows just as for a single observation that

$$E_\theta[\lambda_n'(\boldsymbol{X}|\theta)] = \int_S \cdots \int_S f_n'(\boldsymbol{x}|\theta)\, dx_1 \ldots dx_n = 0.$$

Therefore,

$$\begin{aligned}
\mathrm{Cov}_\theta[T, \lambda_n'(\boldsymbol{X}|\theta)] &= E_\theta[T\lambda_n'(\boldsymbol{X}|\theta)] \\
&= \int_S \cdots \int_S r(\boldsymbol{x})\lambda_n'(\boldsymbol{x}|\theta) f_n(\boldsymbol{x}|\theta)\, dx_1 \ldots dx_n \\
&= \int_S \cdots \int_S r(\boldsymbol{x}) f_n'(\boldsymbol{x}|\theta)\, dx_1 \ldots dx_n.
\end{aligned} \tag{8.8.16}$$

Next, write

$$m(\theta) = \int_S \cdots \int_S r(\boldsymbol{x}) f_n(\boldsymbol{x}|\theta)\, dx_1 \ldots dx_n \quad \text{for } \theta \in \Omega. \tag{8.8.17}$$

Finally, suppose that when both sides of Eq. (8.8.17) are differentiated with respect to θ, the derivative can be taken “inside the integrals” on the left side. Then

$$m'(\theta) = \int_S \cdots \int_S r(\boldsymbol{x}) f_n'(\boldsymbol{x}|\theta)\, dx_1 \ldots dx_n \quad \text{for } \theta \in \Omega. \tag{8.8.18}$$

It follows from Eqs. (8.8.16) and (8.8.18) that

$$\mathrm{Cov}_\theta[T, \lambda_n'(\boldsymbol{X}|\theta)] = m'(\theta) \quad \text{for } \theta \in \Omega. \tag{8.8.19}$$

Theorem 4.6.3 says that

$$\{\mathrm{Cov}_\theta[T, \lambda_n'(\boldsymbol{X}|\theta)]\}^2 \leq \mathrm{Var}_\theta(T)\, \mathrm{Var}_\theta[\lambda_n'(\boldsymbol{X}|\theta)]. \tag{8.8.20}$$

Therefore, it follows from Eqs. (8.8.10), (8.8.12), (8.8.19), and (8.8.20) that Eq. (8.8.14) holds.

Finally, notice that (8.8.14) is an equality if and only if (8.8.20) is an equality. This, in turn, is an equality if and only if there exist nonzero constants a and b and a constant c such that $aT + b\lambda_n(\boldsymbol{X}|\theta) = c$. This last claim follows from the similar statement in Theorem 4.6.3. In all of the calculations concerned with Fisher infomration, we have been treating θ as a constant; hence, the constants a, b, and c just mentioned can depend on θ, but must not depend on $\boldsymbol{X}$. Then $u(\theta) = b/a$ and $v(\theta) = c/a$. ■

The following simple corollary to Theorem 8.8.3 gives a lower bound on the variance of an unbiased estimator of θ.

Corollary 8.8.1 Cramér-Rao Lower Bound on the Variance of an Unbiased Estimator. Assume the assumptions of Theorem 8.8.3. Let T be an unbiased estimator of θ. Then

$$\mathrm{Var}_\theta(T) \ge \frac{1}{nI(\theta)}.$$

Proof Because T is an unbiased estimator of θ, $m(\theta) = \theta$ and $m'(\theta) = 1$ for every value of $\theta \in \Omega$. Now apply Eq. (8.8.14). ■

In words, Corollary 8.8.1 says that the variance of an unbiased estimator of θ cannot be smaller than the reciprocal of the Fisher information in the sample.

Example 8.8.6 Unbiased Estimation of the Parameter of an Exponential Distribution. Let $X_1, \dots, X_n$ be a random sample of size $n > 2$ from the exponential distribution with parameter β. That is, each X_i has p.d.f. $f(x|\beta) = \beta \exp(-\beta x)$ for $x > 0$. Then

$$\lambda(x|\beta) = \log(\beta) - \beta x,$$
$$\lambda'(x|\beta) = \frac{1}{\beta} - x,$$
$$\lambda''(x|\beta) = -\frac{1}{\beta^2}.$$

It can be verified that the conditions required to establish (8.8.3) hold in this example. Then the Fisher information in one observation is

$$I(\beta) = -E_\theta\left[-\frac{1}{\beta^2}\right] = \frac{1}{\beta^2}.$$

The information in the whole sample is then $I_n(\beta) = n/\beta^2$. Consider the estimator $T = (n-1)/\sum_{i=1}^n X_i$. Theorem 5.7.7 says that $\sum_{i=1}^n X_i$ has the gamma distribution with parameters n and β. In Exercise 21 in Sec. 5.7, you proved that the mean and variance of $1/\sum_{i=1}^n X_i$ are $\beta/(n-1)$ and $\beta^2/[(n-1)^2(n-2)]$, respectively. Thus, T is unbiased and its variance is $\beta^2/(n-2)$. The variance is indeed larger than the lower bound, $1/I_n(\beta) = \beta^2/n$. The reason the inequality is strict is that T is not a linear function of $\lambda_n'(\boldsymbol{X}|\theta)$. Indeed, T is 1 over a linear function of $\lambda_n'(\boldsymbol{X}|\theta)$.

On the other hand, if we wish to estimate $m(\beta) = 1/\beta$, $U = \overline{X}_n$ is an unbiased estimator with variance $1/(n\beta^2)$. The information inequality says that the lower bound on the variance of an estimator of $1/\beta$ is

$$\frac{m'(\beta)^2}{n/\beta^2} = \frac{(-1/\beta^2)^2}{n/\beta^2} = \frac{1}{n\beta^2}.$$

In this case, we see that there is equality in (8.8.14). ◄

Example 8.8.7 Studying Customer Arrivals. Return to the store owner in Example 8.8.5 who wants to compare the estimators of θ and $1/\theta$ that she could compute from either the Poisson random variable Y or the gamma random variable X. The case of unbiased estimators based on X was already handled in Example 8.8.6, where our X has the same distribution as $\sum_{i=1}^n X_i$ in that example when $\theta = \beta$. Hence, X/n is an unbiased estimator of $1/\theta$ whose variance equals the Cramér-Rao lower bound, and $(n-1)/X$ is an unbiased estimator of θ whose variance is strictly larger than the lower bound. Since $E_\theta(Y) = t\theta$, we see that Y/t is an unbiased estimator of θ whose variance is also known to be θ/t, which is the Cramér-Rao lower bound. Unfortunately, there is no

unbiased estimator of $1/\theta$ based on Y alone. The estimator $\delta(Y) = t/(Y+1)$ satisfies

$$E_\theta[\delta(Y)] = \frac{1}{\theta}\left[1 - e^{-t\theta}\right].$$

If t is large and θ is not too small, the bias will be small, but it is impossible to find an unbiased estimator. The reason is that the mean of every function of Y is $\exp(-t\theta)$ times a power series in θ. Every such function is differentiable in a neighborhood of $\theta = 0$. The function $1/\theta$ is not differentiable at $\theta = 0$. ◀

Efficient Estimators

Example 8.8.8 Variance of a Poisson Distribution. In Example 8.7.5, we presented a collection of different unbiased estimators of the the variance of a Poisson distribution based on a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from that distribution. After that example, we made the claim that one of the estimators has the smallest variance among the entire collection. The information inequality gives us a way to address comparisons of such collections of estimators without necessarily listing them all or computing their variances. ◀

An estimator whose variance equals the Cramér-Rao lower bound makes the most efficient use of the data $\boldsymbol{X}$ in some sense.

Definition 8.8.3 Efficient Estimator. It is said that an estimator T is an *efficient estimator of its expectation* $m(\theta)$ if there is equality in (8.8.14) for every value of $\theta \in \Omega$.

One difficulty with Definition 8.8.3 is that, in a given problem, there may be no estimator of a particular function $m(\theta)$ whose variance actually attains the Cramér-Rao lower bound. For example, if the random variable X has the normal distribution for which the mean is 0 and the standard deviation σ is unknown ($\sigma > 0$), then it can be shown that the variance of every unbiased estimator of σ based on the single observation X is strictly greater than $1/I(\sigma)$ for every value of $\sigma > 0$ (see Exercise 9). In Example 8.8.6, no efficient estimator of β exists.

On the other hand, in many standard estimation problems there do exist efficient estimators. Of course, the estimator that is identically equal to a constant is an efficient estimator of that constant, since the variance of this estimator is 0. However, as we shall now show, there are often efficient estimators of more interesting functions of θ as well.

According to Theorem 8.8.3, there will be equality in the information inequality (8.8.14) if and only if the estimator T is a linear function of $\lambda_n'(\boldsymbol{X}|\theta)$. It is possible that the only efficient estimators in a given problem will be constants. The reason is as follows: Because T is an estimator, it cannot involve the parameter θ. Therefore, in order for T to be efficient, it must be possible to find functions $u(\theta)$ and $v(\theta)$ such that the parameter θ will actually be canceled from the right side of Eq. (8.8.15), and the value of T will depend only on the observations $\boldsymbol{X}$ and not on θ.

Example 8.8.9 Sampling from a Poisson Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with unknown mean θ ($\theta > 0$). We shall show that $\overline{X}_n$ is an efficient estimator of θ.

The joint p.f. of $X_1, \ldots, X_n$ can be written in the form

$$f_n(\boldsymbol{x}|\theta) = \frac{e^{-n\theta}\theta^{n\overline{x}_n}}{\Pi_{i=I}^n (x_i!)}.$$

Therefore,

$$\lambda_n(\boldsymbol{X}|\theta) = -n\theta + n\overline{X}_n \log\theta - \sum_{i=1}^{n} \log(X_i!)$$

and

$$\lambda_n'(\boldsymbol{X}|\theta) = -n + \frac{n\overline{X}_n}{\theta}. \tag{8.8.21}$$

If we now let $u(\theta) = \theta/n$ and $v(\theta) = \theta$, then it is found from Eq. (8.8.21) that

$$\overline{X}_n = u(\theta)\lambda_n'(\boldsymbol{X}|\theta) + v(\theta).$$

Since the statistic $\overline{X}_n$ has been represented as a linear function of $\lambda_n'(\boldsymbol{X}|\theta)$, it follows that $\overline{X}_n$ is an efficient estimator of its expectation θ. In other words, the variance of $\overline{X}_n$ will attain the lower bound given by the information inequality, which in this example is θ/n (see Exercise 3). This fact can also be verified directly. ◀

Unbiased Estimators with Minimum Variance Suppose that in a given problem a particular estimator T is an efficient estimator of its expectation $m(\theta)$, and let T_1 denote any other unbiased estimator of $m(\theta)$. Then for every value of $\theta \in \Omega$, $\text{Var}_\theta(T)$ will be equal to the lower bound provided by the information inequality, and $\text{Var}_\theta(T_1)$ will be at least as large as that lower bound. Hence, $\text{Var}_\theta(T) \le \text{Var}_\theta(T_1)$ for $\theta \in \Omega$. In other words, if T is an efficient estimator of $m(\theta)$, then among all unbiased estimators of $m(\theta)$, T will have the smallest variance for every possible value of θ.

Example 8.8.10 Variance of a Poisson Distribution. In Example 8.8.9, we saw that $\overline{X}_n$ is an efficient estimator of the mean θ of a Poisson distribution. Therefore, for every value of $\theta > 0$, $\overline{X}_n$ has the smallest variance among all unbiased estimators of θ. Since θ is also the variance of the Poisson distribution with mean θ, we know that $\overline{X}_n$ has the smallest variance among all unbiased estimators of the variance. This establishes the claim that was made without proof after Example 8.7.5. In particular, the estimator $\hat{\sigma}_1^2$ in Example 8.7.5 is not a linear function of $\lambda_n'(\boldsymbol{X}|\theta)$, and hence its variance must be strictly larger than Cramér-Rao lower bound. Similarly, the other estimators in Eq. (8.7.5) must each have variance larger than the Cramér-Rao lower bound. ◀

Properties of Maximum Likelihood Estimators for Large Samples

Suppose that $X_1, \dots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, and suppose also that $f(x|\theta)$ satisfies conditions similar to those which were needed to derive the information inequality. For each sample size n, let $\hat{\theta}_n$ denote the M.L.E. of θ. We shall show that if n is large, then the distribution of $\hat{\theta}_n$ is approximately the normal distribution with mean θ and variance $1/[nI(\theta)]$.

Theorem 8.8.4 Asymptotic Distribution of an Efficient Estimator. Assume the assumptions of Theorem 8.8.3. Let T be an efficient estimator of its mean $m(\theta)$. Assume that $m'(\theta)$ is never 0. Then the asymptotic distribution of

$$\frac{[nI(\theta)]^{1/2}}{m'(\theta)}[T - m(\theta)]$$

is the standard normal distribution.

Proof Consider first the random variable $\lambda'_n(\boldsymbol{X}|\theta)$. Since $\lambda_n(\boldsymbol{X}|\theta) = \sum_{i=1}^n \lambda(X_i|\theta)$, then

$$\lambda'_n(\boldsymbol{X}|\theta) = \sum_{i=1}^{n} \lambda'(X_i|\theta).$$

Furthermore, since the n random variables $X_1, \ldots, X_n$ are i.i.d., the n random variables $\lambda'(X_1|\theta), \ldots, \lambda'(X_n|\theta)$ will also be i.i.d. We know from Eqs. (8.8.7) and (8.8.4) that the mean of each of these variables is 0, and the variance of each is $I(\theta)$. Hence, it follows from the central limit theorem of Lindeberg and Lévy (Theorem 6.3.1) that the asymptotic distribution of the random variable $\lambda'_n(\boldsymbol{X}|\theta)/[nI(\theta)]^{1/2}$ is the standard normal distribution.

Since T is an efficient estimator of $m(\theta)$, we have

$$E_\theta(T) = m(\theta) \quad \text{and} \quad \mathrm{Var}_\theta(T) = \frac{[m'(\theta)]^2}{nI(\theta)}. \tag{8.8.22}$$

Furthermore, there must exist functions $u(\theta)$ and $v(\theta)$ that satisfy Eq. (8.8.15). Because the random variable $\lambda'_n(\boldsymbol{X}|\theta)$ has mean 0 and variance $nI(\theta)$, it follows from Eq. (8.8.15) that

$$E_\theta(T) = v(\theta) \quad \text{and} \quad \mathrm{Var}_\theta(T) = [u(\theta)]^2 nI(\theta).$$

When these values for the mean and the variance of T are compared with the values in Eq. (8.8.22), we find that $v(\theta) = m(\theta)$ and $|u(\theta)| = |m'(\theta)|/[nI(\theta)]$. To be specific, we shall assume that $u(\theta) = m'(\theta)/[nI(\theta)]$, although the same conclusions would be obtained if $u(\theta) = -m'(\theta)/[nI(\theta)]$.

Next, substitute the values $u(\theta) = m'(\theta)/[nI(\theta)]$ and $v(\theta) = m(\theta)$ into Eq. (8.8.15) to obtain

$$T = \frac{m'(\theta)}{nI(\theta)} \lambda'_n(\boldsymbol{X}|\theta) + m(\theta).$$

Rearranging this equation slightly yields

$$\frac{[nI(\theta)]^{1/2}}{m'(\theta)} [T - m(\theta)] = \frac{\lambda'_n(\boldsymbol{X}|\theta)}{[nI(\theta)]^{1/2}}. \tag{8.8.23}$$

We have already shown that the asymptotic distribution of the random variable on the right side of Eq. (8.8.23) is the standard normal distribution. Therefore, the asymptotic distribution of the random variable on the left side of Eq. (8.8.23) is also the standard normal distribution. ∎

Asymptotic Distribution of an M.L.E It follows from Theorem 8.8.4 that if the M.L.E. $\hat{\theta}_n$ is an efficient estimator of θ for each value of n, then the asymptotic distribution of $[nI(\theta)]^{1/2}(\hat{\theta}_n - \theta)$ is the standard normal distribution. However, it can be shown that even in an arbitrary problem in which $\hat{\theta}_n$ is not an efficient estimator, $[nI(\theta)]^{1/2}(\hat{\theta}_n - \theta)$ has this same asymptotic distribution under certain conditions. Without presenting all the required conditions in full detail, we can state the following result. The proof of this result can be found in Schervish (1995, chapter 7).

Theorem 8.8.5 Asymptotic Distribution of M.L.E. Suppose that in an arbitrary problem the M.L.E. $\hat{\theta}_n$ is determined by solving the equation $\lambda'_n(\boldsymbol{x}|\theta) = 0$, and in addition both the second and third derivatives $\lambda''_n(\boldsymbol{x}|\theta)$ and $\lambda'''_n(\boldsymbol{x}|\theta)$ exist and satisfy certain regularity conditions. Then the asymptotic distribution of $[nI(\theta)]^{1/2}(\hat{\theta}_n - \theta)$ is the standard normal distribution. ∎

In practical terms, Theorem 8.8.5 states that in most problems in which the sample size n is large, and the M.L.E. $\hat{\theta}_n$ is found by differentiating the likelihood function $f_n(\boldsymbol{x}|\theta)$ or its logarithm, the distribution of $[nI(\theta)]^{1/2}(\hat{\theta}_n - \theta)$ will be approximately the standard normal distribution. Equivalently, the distribution of $\hat{\theta}_n$ will be approximately the normal distribution with mean θ and variance $1/[nI(\theta)]$. Under these conditions, it is said that $\hat{\theta}_n$ is an *asymptotically efficient estimator*.

Example 8.8.11 Estimating the Standard Deviation of a Normal Distribution. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with known mean 0 and unknown standard deviation σ ($\sigma > 0$). It can be shown that the M.L.E. of σ is

$$\hat{\sigma} = \left[\frac{1}{n} \sum_{i=1}^{n} X_i^2 \right]^{1/2}.$$

Also, it can be shown (see Exercise 4) that the Fisher information in a single observation is $I(\sigma) = 2/\sigma^2$. Therefore, if the sample size n is large, the distribution of $\hat{\sigma}$ will be approximately the normal distribution with mean σ and variance $\sigma^2/(2n)$. ◀

For cases in which it is difficult to compute the M.L.E., there is a result similar to Theorem 8.8.5. The proof of Theorem 8.8.6 can also be found as a special case of theorem 7.75 in Schervish (1995).

Theorem 8.8.6 Efficient Estimation. Assume the same smoothness conditions on the likelihood function as in Theorem 8.8.5. Assume that $\tilde{\theta}_n$ is a sequence of estimators of θ such that $\sqrt{n}(\tilde{\theta}_n - \theta)$ converges in distribution to some distribution (it doesn't matter what distribution). Use $\tilde{\theta}_n$ as the starting value, and perform one step of Newton's method (Definition 7.6.2) toward finding the M.L.E. of θ. Let the result of this one step be called θ_n^*. Then the asymptotic distribution of $[nI(\theta)]^{1/2}(\theta_n^* - \theta)$ is the standard normal distribution. ■

A typical choice of $\tilde{\theta}_n$ in Theorem 8.8.6 is a method of moments estimator (Definition 7.6.3). Example 7.6.6 illustrates such an application of Theorem 8.8.6 when sampling from a gamma distribution.

The Bayesian Point of View Another general property of the M.L.E. $\hat{\theta}_n$ pertains to making inferences about a parameter θ from the Bayesian point of view. Suppose that the prior distribution of θ is represented by a positive and differentiable p.d.f. over the interval Ω, and the sample size n is large. Then under conditions similar to the regularity conditions that are needed to assure the asymptotic normality of the distribution of $\hat{\theta}_n$, it can be shown that the posterior distribution of θ, after the values of $X_1, \dots, X_n$ have been observed, will be approximately the normal distribution with mean $\hat{\theta}_n$ and variance $1/[nI(\hat{\theta}_n)]$.

Example 8.8.12 The Posterior Distribution of the Standard Deviation. Suppose again that $X_1, \dots, X_n$ form a random sample from the normal distribution with known mean 0 and unknown standard deviation σ. Suppose also that the prior p.d.f. of σ is a positive and differentiable function for $\sigma > 0$, and the sample size n is large. Since $I(\sigma) = 2/\sigma^2$, it follows that the posterior distribution of σ will be approximately the normal distribution with mean $\hat{\sigma}$ and variance $\hat{\sigma}^2/(2n)$, where $\hat{\sigma}$ is the M.L.E. of σ calculated from the observed values in the sample. Figure 8.9 illustrates this approximation based on a

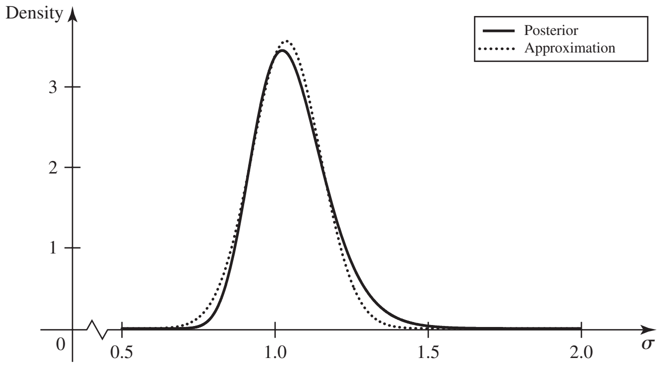


Figure 8.9 Posterior p.d.f. of σ and approximation based on Fisher information in Example 8.8.12.

sample of $n = 40$ i.i.d. simulated normal random variables with mean 0 and variance 1. In this sample, the M.L.E. was $\hat{\sigma} = 1.061$. Figure 8.9 shows the actual posterior p.d.f. based on an improper prior with "p.d.f." $1/\sigma$ together with the approximate normal posterior p.d.f. with mean 1.061 and variance $1.061^2/80 = 0.0141$. ◀

Fisher Information for Multiple Parameters

Example 8.8.13 **Sample from a Normal Distribution.** Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the normal distribution with mean μ and variance σ^2. Is there an analog to Fisher information for the vector parameter $\theta = (\mu, \sigma^2)$? ◀

In the spirit of Definition 8.8.1 and Theorem 8.8.1, we define Fisher information in terms of derivatives of the logarithm of the likelihood function. We shall define the Fisher information in a random sample of size n with the understanding that the Fisher information in a single random variable corresponds to a sample size of $n = 1$.

Definition 8.8.4 **Fisher Information for a Vector Parameter.** Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from a distribution for which the p.d.f. is $f(x|\theta)$, where the value of the parameter $\theta = (\theta_1, \ldots, \theta_k)$ must lie in an open subset Ω of a k-dimensional real space. Let $f_n(\boldsymbol{x}|\theta)$ denote the joint p.d.f. or joint p.f. of $\boldsymbol{X}$. Define

$$\lambda_n(\boldsymbol{x}|\theta) = \log f_n(x|\theta).$$

Assume that the set of $\boldsymbol{x}$ such that $f_n(\boldsymbol{x}|\theta) > 0$ is the same for all θ and that $\log f_n(\boldsymbol{x}|\theta)$ is twice differentiable with respect to θ. The *Fisher information matrix* $I_n(\theta)$ in the random sample $\boldsymbol{X}$ is defined as the $k \times k$ matrix with (i, j) element equal to

$$I_{n,i,j}(\theta) = \text{Cov}_\theta\left[\frac{\partial}{\partial\theta_i}\lambda'_n(\boldsymbol{X}|\theta), \frac{\partial}{\partial\theta_j}\lambda'_n(\boldsymbol{X}|\theta)\right].$$

Example 8.8.14 **Sample from a Normal Distribution.** In Example 8.8.13, let $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. As in Eq. (7.5.3), we obtain

$$\lambda_n(\boldsymbol{X}|\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\theta_2) - \frac{1}{2\theta_2}\sum_{i=1}^{n}(X_i - \theta_1)^2.$$

The first partial derivatives are

$$\frac{\partial}{\partial\theta_1}\lambda_n(\boldsymbol{x}|\theta) = \frac{1}{\theta_2}\sum_{i=1}^{n}(X_i - \theta_1), \tag{8.8.24}$$

$$\frac{\partial}{\partial\theta_2}\lambda_n(\boldsymbol{x}|\theta) = \frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(X_i - \theta_1)^2. \tag{8.8.25}$$

Since the means of the two random variables above are both 0, their covariances are the means of the products. The distribution of $\sum_{i=1}^{n}(X_i - \theta_1)$ is the normal distribution with mean 0 and variance $n\theta_2$. The distribution of $\sum_{i=1}^{n}(X_i - \theta_1)^2/\theta_2$ is the χ^2 distribution with n degrees of freedom. So the variance of (8.8.24) is n/θ_2, and the variance of (8.8.25) is $2n/\theta_2^2$. The mean of the product of (8.8.24) and (8.8.25) is 0 because the third central moment of a normal distribution is 0. This makes

$$I_n(\theta) = \begin{pmatrix} \frac{n}{\theta_2} & 0 \\ 0 & \frac{n}{\theta_2^2} \end{pmatrix}. \quad ◀$$

The results for one-dimensional parameters all have versions for k-dimensional parameters. For example, in Eq. (8.8.3), $\lambda''(X|\theta)$ is replaced by the $k \times k$ matrix of second partial derivatives. In the Cramér-Rao inequality, we need the inverse of the matrix $I_n(\theta)$, and $m'(\theta)$ must be replaced by the vector of partial derivatives. Specifically, if T is a statistic with finite variance and mean $m(\theta)$, then

$$\mathrm{Var}_\theta(T) \geq \left(\frac{\partial}{\partial\theta_1}m(\theta), \ldots, \frac{\partial}{\partial\theta_k}m(\theta)\right) I_n(\theta)^{-1} \begin{pmatrix} \frac{\partial}{\partial\theta_1}m(\theta) \\ \vdots \\ \frac{\partial}{\partial\theta_k}m(\theta) \end{pmatrix}. \tag{8.8.26}$$

Also, the inequality in (8.8.26) is equality if and only if T is a linear function of the vector

$$\left(\frac{\partial}{\partial\theta_1}\lambda_n(\boldsymbol{x}|\theta), \ldots, \frac{\partial}{\partial\theta_k}\lambda_n(\boldsymbol{x}|\theta)\right). \tag{8.8.27}$$

Example 8.8.15

Sample from a Normal Distribution. In Example 8.8.14, the coordinates of the vector in (8.8.27) are linear functions of the two random variables $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$. So, the only statistics whose variances equal the lower bound in (8.8.26) are of the form $T = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 + c$. The mean of such a statistic T is

$$E_\theta(T) = an\theta_1 + bn(\theta_2 + \theta_1^2) + c. \tag{8.8.28}$$

In particular, it is impossible to obtain θ_2 as a special case of (8.8.28). There is no efficient unbiased estimator of $\theta_2 = \sigma^2$. It can be proven that $(\sigma')^2$, which was defined in Eq. (8.4.3), is an unbiased estimator that has minimum variance among all unbiased estimators. The proof of this fact is beyond the scope of this text. The variance of $(\sigma')^2$ is $2\theta_2^2/(n-1)$, while the Cramér-Rao lower bound is $2\theta_2^2/n$. ◀

Example 8.8.16

Multinomial Distributions. Let $\boldsymbol{X} = (X_1, \ldots, X_k)$ have the multinomial distribution with parameters n and $\boldsymbol{p} = (p_1, \ldots, p_k)$ as defined in Definition 5.9.1. Finding the Fisher information in this example involves a subtle point. The parameter vector $\boldsymbol{p}$ takes values in the set

$$\{\boldsymbol{p} : p_1 + \cdots + p_k = 1, \text{ all } p_i \geq 0\}.$$

No subset of this set is open. Hence, no matter what set we choose for the parameter space, Definition 8.8.4 does not apply to this parameter. However, there is an

equivalent paramter $\boldsymbol{p}^* = (p_1, \ldots, p_{k-1})$ that takes values in the set

$$\{\boldsymbol{p}^* : p_1 + \cdots + p_{k-1} \le 1, \text{ all } p_i \ge 0\},$$

which has nonempty interior. With this version of the parameter, and assuming that the parameter space is the interior of the set above, it is straightforward to calculate the Fisher information, as in Exercise 20. ◄

Summary

Fisher information attempts to measure the amount of information about a parameter that a random variable or sample contains. Fisher information from independent random variables adds together to form the Fisher information in the sample. The information inequality (Cramér-Rao lower bound) provides lower bounds on the variances of all estimators. An estimator is efficient if its variance equals the lower bound. The asymptotic distribution of a maximum likelihood estimator of θ is (under regularity conditions) normal with mean θ and variance equal to 1 over the Fisher information in the sample. Also, for large sample sizes, the posterior distribution of θ is approximately normal with mean equal to the M.L.E. and variance equal to 1 over the Fisher information in the sample evaluated at the M.L.E.

Exercises

1. Suppose that a random variable X has a normal distribution for which the mean μ is unknown ($-\infty < \mu < \infty$) and the variance σ^2 is known. Let $f(x|\mu)$ denote the p.d.f. of X, and let $f'(x|\mu)$ and $f''(x|\mu)$ denote the first and second partial derivatives with respect to μ. Show that

$$\int_{-\infty}^{\infty} f'(x|\mu)\,dx = 0 \quad \text{and} \quad \int_{-\infty}^{\infty} f''(x|\mu)\,dx = 0.$$

2. Suppose that X has the geometric distribution with parameter p. (See Sec. 5.5.) Find the Fisher information $I(p)$ in X.

3. Suppose that a random variable X has the Poisson distribution with unknown mean $\theta > 0$. Find the Fisher information $I(\theta)$ in X.

4. Suppose that a random variable has the normal distribution with mean 0 and unknown standard deviation $\sigma > 0$. Find the Fisher information $I(\sigma)$ in X.

5. Suppose that a random variable X has the normal distribution with mean 0 and unknown variance $\sigma^2 > 0$. Find the Fisher information $I(\sigma^2)$ in X. Note that in this exercise the variance σ^2 is regarded as the parameter, whereas in Exercise 4 the standard deviation σ is regarded as the parameter.

6. Suppose that X is a random variable for which the p.d.f. or the p.f. is $f(x|\theta)$, where the value of the parameter θ is unknown but must lie in an open interval Ω. Let $I_0(\theta)$ denote the Fisher information in X. Suppose now that the parameter θ is replaced by a new parameter μ, where $\theta = \psi(\mu)$, and ψ is a differentiable function. Let $I_1(\mu)$ denote the Fisher information in X when the parameter is regarded as μ. Show that

$$I_1(\mu) = [\psi'(\mu)]^2 I_0[\psi(\mu)].$$

7. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p. Show that $\overline{X}_n$ is an efficient estimator of p.

8. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance $\sigma^2 > 0$. Show that $\overline{X}_n$ is an efficient estimator of μ.

9. Suppose that a single observation X is taken from the normal distribution with mean 0 and unknown standard deviation $\sigma > 0$. Find an unbiased estimator of σ, determine its variance, and show that this variance is greater than $1/I(\sigma)$ for every value of $\sigma > 0$. Note that the value of $I(\sigma)$ was found in Exercise 4.

10. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean 0 and unknown standard deviation $\sigma > 0$. Find the lower bound specified by the information inequality for the variance of any unbiased estimator of $\log \sigma$.

11. Suppose that $X_1, \ldots, X_n$ form a random sample from an exponential family for which the p.d.f. or the p.f. $f(x|\theta)$ is as specified in Exercise 23 of Sec. 7.3. Suppose also that the unknown value of θ must belong to an open interval Ω of the real line. Show that the estimator $T = \sum_{i=1}^{n} d(X_i)$ is an efficient estimator. *Hint:* Show that T can be represented in the form given in Eq. (8.8.15).

12. Suppose that $X_1, \dots, X_n$ form a random sample from a normal distribution for which the mean is known and the variance is unknown. Construct an efficient estimator that is not identically equal to a constant, and determine the expectation and the variance of this estimator.

13. Determine what is wrong with the following argument: Suppose that the random variable X has the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown ($\theta > 0$). Then $f(x|\theta) = 1/\theta$, $\lambda(x|\theta) = -\log\theta$ and $\lambda'(x|\theta) = -(1/\theta)$. Therefore,

$$I(\theta) = E_\theta\{[\lambda'(X|\theta)]^2\} = \frac{1}{\theta^2}.$$

Since $2X$ is an unbiased estimator of θ, the information inequality states that

$$\mathrm{Var}(2X) \ge \frac{1}{I(\theta)} = \theta^2.$$

But

$$\mathrm{Var}(2X) = 4\,\mathrm{Var}(X) = 4 \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3} < \theta^2.$$

Hence, the information inequality is not correct.

14. Suppose that $X_1, \dots, X_n$ form a random sample from the gamma distribution with parameters α and β, where α is unknown and β is known. Show that if n is large, the distribution of the M.L.E. of α will be approximately a normal distribution with mean α and variance

$$\frac{[\Gamma(\alpha)]^2}{n\{\Gamma(\alpha)\Gamma''(\alpha) - [\Gamma'(\alpha)]^2\}}.$$

15. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with uknown mean μ and known variance σ^2, and the prior p.d.f. of μ is a positive and differentiable function over the entire real line. Show that if n is large, the posterior distribution of μ given that $X_i = x_i$ ($i = 1, \dots, n$) will be approximately a normal distribution with mean $\overline{x}_n$ and variance σ^2/n.

16. Suppose that $X_1, \dots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p, and the prior p.d.f. of p is a positive and differentiable function over the interval $0 < p < 1$. Suppose, furthermore, that n is large, the observed values of $X_1, \dots, X_n$ are $x_1, \dots, x_n$, and $0 < \overline{x}_n < 1$. Show that the posterior distribution of p will be approximately a normal distribution with mean $\overline{x}_n$ and variance $\overline{x}_n(1 - \overline{x}_n)/n$.

17. Let X have the binomial distribution with parameters n and p. Assume that n is known. Show that the Fisher information in X is $I(p) = n/[p(1-p)]$.

18. Let X have the negative binomial distribution with parameters r and p. Assume that r is known. Show that the Fisher information in X is $I(p) = r/[p^2(1-p)]$.

19. Let X have the gamma distribution with parameters n and θ with θ unknown. Show that the Fisher information in X is $I(\theta) = n/\theta^2$.

20. Find the Fisher information matrix about $\boldsymbol{p}^*$ in Example 8.8.16.

8.9 Supplementary Exercises

1. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with known mean 0 and unknown variance σ^2. Show that $\sum_{i=1}^n X_i^2/n$ is the unbiased estimator of σ^2 that has the smallest possible variance for all possible values of σ^2.

2. Prove that if X has the t distribution with one degree of freedom, then $1/X$ also has the t distribution with one degree of freedom.

3. Suppose that U and V are independent random variables, and that each has the standard normal distribution. Show that U/V, $U/|V|$, and $|U|/V$ each has the t distribution with one degree of freedom.

4. Suppose that X_1 and X_2 are independent random variables, and that each has the normal distribution with mean 0 and variance σ^2. Show that $(X_1 + X_2)/(X_1 - X_2)$ has the t distribution with one degree of freedom.

5. Suppose that $X_1, \dots, X_n$ form a random sample from the exponential distribution with parameter β. Show that $2\beta\sum_{i=1}^n X_i$ has the χ^2 distribution with $2n$ degrees of freedom.

6. Suppose that $X_1, \dots, X_n$ form a random sample from an unknown probability distribution P on the real line. Let A be a given subset of the real line, and let $\theta = P(A)$. Construct an unbiased estimator of θ, and specify its variance.

7. Suppose that $X_1, \dots, X_m$ form a random sample from the normal distribution with mean μ_1 and variance σ^2, and $Y_1, \dots, Y_n$ form an independent random sample from the normal distribution with mean μ_2 and variance $2\sigma^2$. Let $S_X^2 = \sum_{i=1}^m (X_i - \overline{X}_m)^2$ and $S_Y^2 = \sum_{i=1}^n (Y_i - \overline{Y}_n)^2$.

a. For what pairs of values of α and β is $\alpha S_X^2 + \beta S_Y^2$ an unbiased estimator of σ^2?

b. Determine the values of α and β for which $\alpha S_X^2 + \beta S_Y^2$ will be an unbiased estimator with minimum variance.

8. Suppose that $X_1, \dots, X_{n+1}$ form a random sample from a normal distribution, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $T_n = \left[\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right]^{1/2}$. Determine the value of a constant k such that the random variable $k(X_{n+1} - \overline{X}_n)/T_n$ will have a t distribution.

9. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2, and Y is an independent random variable having the normal distribution with mean 0 and variance $4\sigma^2$. Determine a function of $X_1, \dots, X_n$ and Y that does not involve μ or σ^2 but has the t distribution with $n-1$ degrees of freedom.

10. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2, where both μ and σ^2 are unknown. A confidence interval for μ is to be constructed with confidence coefficient 0.90. Determine the smallest value of n such that the expected squared length of this interval will be less than $\sigma^2/2$.

11. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Construct a lower confidence limit $L(X_1, \dots, X_n)$ for μ such that

$$\Pr[\mu > L(X_1, \dots, X_n)] = 0.99.$$

12. Consider again the conditions of Exercise 11. Construct an upper confidence limit $U(X_1, \dots, X_n)$ for σ^2 such that

$$\Pr[\sigma^2 < U(X_1, \dots, X_n)] = 0.99.$$

13. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean θ and known variance σ^2. Suppose also that the prior distribution of θ is normal with mean μ and variance ν^2.

a. Determine the shortest interval I such that $\Pr(\theta \in I | x_1, \dots, x_n) = 0.95$, where the probability is calculated with respect to the posterior distribution of θ, as indicated.

b. Show that as $\nu^2 \to \infty$, the interval I converges to an interval I^* that is a confidence interval for θ with confidence coefficient 0.95.

14. Suppose that $X_1, \dots, X_n$ form a random sample from the Poisson distribution with unknown mean θ, and let $Y = \sum_{i=1}^n X_i$.

a. Determine the value of a constant c such that the estimator e^{-cY} is an unbiased estimator of $e^{-\theta}$.

b. Use the information inequality to obtain a lower bound for the variance of the unbiased estimator found in part (a).

15. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution for which the p.d.f. is as follows:

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

where the value of θ is unknown ($\theta > 0$). Determine the asymptotic distribution of the M.L.E. of θ. (*Note*: The M.L.E. was found in Exercise 9 of Sec. 7.5.)

16. Suppose that a random variable X has the exponential distribution with mean θ, which is unknown ($\theta > 0$). Find the Fisher information $I(\theta)$ in X.

17. Suppose that $X_1, \dots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p. Show that the variance of every unbiased estimator of $(1-p)^2$ must be at least $4p(1-p)^3/n$.

18. Suppose that $X_1, \dots, X_n$ form a random sample from the exponential distribution with unknown parameter β. Construct an efficient estimator that is not identically equal to a constant, and determine the expectation and the variance of this estimator.

19. Suppose that $X_1, \dots, X_n$ form a random sample from the exponential distribution with unknown parameter β. Show that if n is large, the distribution of the M.L.E. of β will be approximately a normal distribution with mean β and variance β^2/n.

20. Consider again the conditions of Exercise 19, and let $\hat{\beta}_n$ denote the M.L.E. of β.

a. Use the delta method to determine the asymptotic distribution of $1/\hat{\beta}_n$.

b. Show that $1/\hat{\beta}_n = \overline{X}_n$, and use the central limit theorem to determine the asymptotic distribution of $1/\hat{\beta}_n$.

21. Let $X_1, \dots, X_n$ be a random sample from the Poisson distribution with mean θ. Let $Y = \sum_{i=1}^n X_i$.

a. Prove that there is no unbiased estimator of $1/\theta$. (*Hint:* Write the equation that is equivalent to $E_\theta(r(X)) = 1/\theta$. Simplify it, and then use what you know from calculus of infinite series to show that no function r can satisfy the equation.)

b. Suppose that we wish to estimate $1/\theta$. Consider $r(Y) = n/(Y+1)$ as an estimator of θ. Find the bias of $r(Y)$, and show that the bias goes to 0 as $n \to \infty$.

c. Use the delta method to find the asymptotic (as $n \to \infty$) distribution of $n/(Y+1)$.

22. Let $X_1, \dots, X_n$ be conditionally i.i.d. with the uniform distribution on the interval $[0, \theta]$. Let $Y_n = \max\{X_1, \dots, X_n\}$.

a. Find the p.d.f. and the quantile function of Y_n/θ.

b. Y_n is often used as an estimator of θ even though it has bias. Compute the bias of Y_n as an estimator of θ.

c. Prove that Y_n/θ is a pivotal.

d. Find a confidence interval for θ with coefficient γ.

Testing Hypotheses

Chapter 9

9.1 Problems of Testing Hypotheses

In Example 8.3.1 on page 473, we were interested in whether or not the mean log-rainfall μ from seeded clouds was greater than some constant, specifically 4. Hypothesis testing problems are similar in nature to the decision problem of Example 8.3.1. In general, hypothesis testing concerns trying to decide whether a parameter θ lies in one subset of the parameter space or in its complement. When θ is one-dimensional, at least one of the two subsets will typically be an interval, possibly degenerate. In this section, we introduce the notation and some common methodology associated with hypothesis testing. We also demonstrate an equivalence between hypothesis tests and confidence intervals.

The Null and Alternative Hypotheses

Example 9.1.1 Rain from Seeded Clouds. In Example 8.3.1, we modeled the log-rainfalls from 26 seeded clouds as normal random variables with unknown mean μ and unknown variance σ^2. Let $\theta = (\mu, \sigma^2)$ denote the parameter vector. We are interested in whether or not $\mu > 4$. To word this in terms of the parameter vector, we are interested in whether or not θ lies in the set $\{(\mu, \sigma^2) : \mu > 4\}$. In Example 8.6.4, we calculated the probability that $\mu > 4$ as part of a Bayesian analysis. If one does not wish to do a Bayesian analysis, one must address the question of whether or not $\mu > 4$ by other means, such as those introduced in this chapter. ◄

Consider a statistical problem involving a parameter θ whose value is unknown but must lie in a certain parameter space Ω. Suppose now that Ω can be partitioned into two disjoint subsets Ω_0 and Ω_1, and the statistician is interested in whether θ lies in Ω_0 or in Ω_1.

We shall let H_0 denote the hypothesis that $\theta \in \Omega_0$ and let H_1 denote the hypothesis that $\theta \in \Omega_1$. Since the subsets Ω_0 and Ω_1 are disjoint and $\Omega_0 \cup \Omega_1 = \Omega$, exactly one of the hypotheses H_0 and H_1 must be true. The statistician must decide which of the hypotheses H_0 or H_1 appears to be true. A problem of this type, in which there are only two possible decisions, is called a problem of *testing hypotheses*. If the statistician makes the wrong decision, he might suffer a certain loss or pay a certain cost. In many problems, he will have an opportunity to observe some data before he has to make his

decision, and the observed values will provide him with information about the value of θ. A procedure for deciding which hypothesis to choose is called a *test procedure* or simply a *test*.

In our discussion up to this point, we have treated the hypotheses H_0 and H_1 on an equal basis. In most problems, however, the two hypotheses are treated quite differently.

Definition 9.1.1 Null and Alternative Hypotheses/Reject. The hypothesis H_0 is called the *null hypothesis* and the hypothesis H_1 is called the *alternative hypothesis*. When performing a test, if we decide that θ lies in Ω_1, we are said to *reject* H_0. If we decide that θ lies in Ω_0, we are said not to reject H_0.

The terminology referring to the decisions in Definition 9.1.1 is asymmetric with regard to the null and alternative hypotheses. We shall return to this point later in the section.

Example 9.1.2 Egyptian Skulls. Manly (1986, p.4) reports measurements of various dimensions of human skulls found in Egypt from various time periods. These data are attributed to Thomson and Randall-Maciver (1905). One time period is approximately 4000 B.C. We might model the observed breadth measurements (in mm) of the skulls as normal random variables with unknown mean μ and variance 26. Interest might lie in how μ compares to the breadth of a modern-day skull, about 140mm. The parameter space Ω could be the positive numbers, and we could let Ω_0 be the interval $[140, \infty)$ while $\Omega_1 = (0, 140)$. In this case, we would write the null and alternative hypotheses as

$$\begin{aligned} H_0&: \quad \mu \geq 140, \\ H_1&: \quad \mu < 140. \end{aligned}$$

More realistically, we would assume that both the mean and variance of breadth measurements were unknown. That is, each measurement is a normal random variable with mean μ and variance σ^2. In this case, the parameter would be two-dimensional, for example, $\theta = (\mu, \sigma^2)$. The parameter space Ω would then be pairs of real numbers. In this case, $\Omega_0 = [140, \infty) \times (0, \infty)$ and $\Omega_1 = (0, 140) \times (0, \infty)$, since the hypotheses only concern the first coordinate μ. The hypotheses to be tested are the same as above, but now μ is only one coordinate of a two-dimensional parameter vector. We will address problems of this type in Sec. 9.5. ◀

How did we decide that the null hypothesis should be $H_0: \mu \geq 140$ in Example 9.1.2 rather than $\mu \leq 140$? Would we be led to the same conclusion either way? We can address these issues after we introduce the possible errors that can arise in hypothesis testing (Definition 9.1.7).

Simple and Composite Hypotheses

Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where the value of the parameter θ must lie in the parameter space Ω; Ω_0 and Ω_1 are disjoint sets with $\Omega_0 \cup \Omega_1 = \Omega$; and it is desired to test the following hypotheses:

$$\begin{aligned} H_0&: \quad \theta \in \Omega_0, \\ H_1&: \quad \theta \in \Omega_1. \end{aligned}$$

For $i = 0$ or $i = 1$, the set Ω_i may contain just a single value of θ or it might be a larger set.

Definition 9.1.2 Simple and Composite Hypotheses. If Ω_i contains just a single value of θ, then H_i is a *simple hypothesis*. If the set Ω_i contains more than one value of θ, then H_i is a *composite hypothesis*.

Under a simple hypothesis, the distribution of the observations is completely specified. Under a composite hypothesis, it is specified only that the distribution of the observations belongs to a certain class. For example, a simple null hypothesis H_0 must have the form

$$H_0: \quad \theta = \theta_0. \tag{9.1.1}$$

Definition 9.1.3 One-Sided and Two-Sided Hypotheses. Let θ be a one-dimensional parameter. *One-sided* null hypotheses are of the form $H_0 : \theta \le \theta_0$ or $H_0 : \theta \ge \theta_0$, with the corresponding one-sided alternative hypotheses being $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$. When the null hypothesis is simple, such as (9.1.1), the alternative hypothesis is usually *two-sided*, $H_1 : \theta \ne \theta_0$.

The hypotheses in Example 9.1.2 are one-sided. In Example 9.1.3 (coming up shortly), the alternative hypothesis is two-sided. One-sided and two-sided hypotheses will be discussed in more detail in Sections 9.3 and 9.4.

The Critical Region and Test Statistics

Example 9.1.3 Testing Hypotheses about the Mean of a Normal Distribution with Known Variance. Suppose that $\boldsymbol{X} = (X_1, \dots, X_n)$ is a random sample from the normal distribution with unknown mean μ and known variance σ^2. We wish to test the hypotheses

$$\begin{aligned} H_0&: \quad \mu = \mu_0, \\ H_1&: \quad \mu \ne \mu_0. \end{aligned} \tag{9.1.2}$$

It might seem reasonable to reject H_0 if $\overline{X}_n$ is far from μ_0. For example, we could choose a number c and reject H_0 if the distance from $\overline{X}_n$ to μ_0 is more than c. One way to express this is by dividing the set S of all possible data vectors $\boldsymbol{x} = (x_1, \dots, x_n)$ (the sample space) into the two sets

$$S_0 = \{\boldsymbol{x} : -c \le \overline{X}_n - \mu_0 \le c\}, \quad \text{and} \quad S_1 = S_0^C.$$

We then reject H_0 if $\boldsymbol{X} \in S_1$, and we don't reject H_0 if $\boldsymbol{X} \in S_0$. A simpler way to express the procedure is to define the statistic $T = |\overline{X}_n - \mu_0|$, and reject H_0 if $T \ge c$. ◀

In general, consider a problem in which we wish to test the following hypotheses:

$$H_0: \theta \in \Omega_0, \quad \text{and} \quad H_1: \theta \in \Omega_1. \tag{9.1.3}$$

Suppose that before the statistician has to decide which hypothesis to choose, she can observe a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ drawn from a distribution that involves the unknown parameter θ. We shall let S denote the sample space of the n-dimensional random vector $\boldsymbol{X}$. In other words, S is the set of all possible values of the random sample.

In a problem of this type, the statistician can specify a test procedure by partitioning the sample space S into two subsets. One subset S_1 contains the values of $\boldsymbol{X}$ for which she will reject H_0, and the other subset S_0 contains the values of $\boldsymbol{X}$ for which she will not reject H_0.

Definition 9.1.4 Critical Region. The set S_1 defined above is called the *critical region* of the test.

In summary, a test procedure is determined by specifying the critical region of the test. The complement of the critical region must then contain all the outcomes for which H_0 will not be rejected.

In most hypothesis-testing problems, the critical region is defined in terms of a statistic, $T = r(\boldsymbol{X})$.

Definition 9.1.5 Test Statistic/Rejection Region. Let $\boldsymbol{X}$ be a random sample from a distribution that depends on a parameter θ. Let $T = r(\boldsymbol{X})$ be a statistic, and let R be a subset of the real line. Suppose that a test procedure for the hypotheses (9.1.3) is of the form "reject H_0 if $T \in R$." Then we call T a *test statistic*, and we call R the *rejection region* of the test.

When a test is defined in terms of a test statistic T and rejection region R, as in Definition 9.1.5, the set $S_1 = \{\boldsymbol{x} : r(\boldsymbol{x}) \in R\}$ is the critical region from Definition 9.1.4.

Typically, the rejection region for a test based on a test statistic T will be some fixed interval or the outside of some fixed interval. For example, if the test rejects H_0 when $T \geq c$, the rejection region is the interval $[c, \infty)$. Once a test statistic is being used, it is simpler to express everything in terms of the test statistic rather than try to compute the critical region from Definition 9.1.4. All of the tests in the rest of this book will be based on test statistics. Indeed, most of the tests can be written in the form "reject H_0 if $T \geq c$." (Example 9.1.7 is one of the rare exceptions.)

In Example 9.1.3, the test statistic is $T = |\overline{X}_n - \mu_0|$, and the rejection region is the interval $[c, \infty)$. One can choose a test statistic using intuitive criteria, as in Example 9.1.3, or based on theoretical considerations. Some theoretical arguments are given in Sections 9.2–9.4 for choosing certain test statistics in a variety of problems involving a single parameter. Although these theoretical results provide optimal tests in the situations in which they apply, many practical problems do not satisfy the conditions required to apply these results.

Example 9.1.4 Rain from Seeded Clouds. We can formulate the problem described in Example 9.1.1 as that of testing the hypotheses $H_0 : \mu \leq 4$ versus $H_1 : \mu > 4$. We could use the same test statistic as in Example 9.1.3. Alternatively, we could use the statistic $U = n^{1/2}(\overline{X}_n - 4)/\sigma'$, which looks a lot like the random variable from Eq. (8.5.1) on which confidence intervals were based. It makes sense, in this case, to reject H_0 if U is large, since that would correspond to $\overline{X}_n$ being large compared to 4. ◀

Note: Dividing Both Parameter Space and Sample Space. In the various definitions given so far, the reader needs to keep straight two different divisions. First, we divided the parameter space Ω into two disjoint subsets, Ω_0 and Ω_1. Next, we divided the sample space S into two disjoint subsets S_0 and S_1. These divisions are related to each other, but they are not the same. For one thing, the parameter space and the sample space usually are of different dimensions, so Ω_0 will necessarily be different from S_0. The relation between the two divisions is the following: If the random sample $\boldsymbol{X}$ lies in the critical region S_1, then we reject the null hypothesis Ω_0. If $\boldsymbol{X} \in S_0$, we don't reject Ω_0. We eventually learn which set S_0 or S_1 contains $\boldsymbol{X}$. We rarely learn which set Ω_0 or Ω_1 contains θ.

The Power Function and Types of Error

Let δ stand for a test procedure of the form discussed earlier in this section, either based on a critical region or based on a test statistic. The interesting probabilistic

properties of δ can be summarized by computing, for each value of $\theta \in \Omega$, either the probability $\pi(\theta|\delta)$ that the test δ will reject H_0 or the probability $1 - \pi(\theta|\delta)$ that it does not reject H_0.

Definition 9.1.6 Power Function. Let δ be a test procedure. The function $\pi(\theta|\delta)$ is called the *power function* of the test δ. If S_1 denotes the critical region of δ, then the power function $\pi(\theta|\delta)$ is determined by the relation

$$\pi(\theta|\delta) = \Pr(\boldsymbol{X} \in S_1|\theta) \quad \text{for} \quad \theta \in \Omega. \tag{9.1.4}$$

If δ is described in terms of a test statistic T and rejection region R, the power function is

$$\pi(\theta|\delta) = \Pr(T \in R|\theta) \quad \text{for} \quad \theta \in \Omega. \tag{9.1.5}$$

Since the power function $\pi(\theta|\delta)$ specifies, for each possible value of the parameter θ, the probability that δ will reject H_0, it follows that the ideal power function would be one for which $\pi(\theta|\delta) = 0$ for every value of $\theta \in \Omega_0$ and $\pi(\theta|\delta) = 1$ for every value of $\theta \in \Omega_1$. If the power function of a test δ actually had these values, then regardless of the actual value of θ, δ would lead to the correct decision with probability 1. In a practical problem, however, there would seldom exist any test procedure having this ideal power function.

Example 9.1.5 Testing Hypotheses about the Mean of a Normal Distribution with Known Variance. In Example 9.1.3, the test δ is based on the test statistic $T = |\overline{X}_n - \mu_0|$ with rejection region $R = [c, \infty)$. The distribution of $\overline{X}_n$ is the normal distribution with mean μ and variance σ^2/n. The parameter is μ because we have assumed that σ^2 is known. The power function can be computed from this distribution. Let Φ denote the standard normal c.d.f. Then

$$\begin{aligned}\Pr(T \in R|\mu) &= \Pr(\overline{X}_n \geq \mu_0 + c|\mu) + \Pr(\overline{X}_n \leq \mu_0 - c|\mu)\\ &= 1 - \Phi\left(n^{1/2}\frac{\mu_0 + c - \mu}{\sigma}\right) + \Phi\left(n^{1/2}\frac{\mu_0 - c - \mu}{\sigma}\right).\end{aligned}$$

The final expression above is the power function $\pi(\mu|\delta)$. Figure 9.1 plots the power functions of three different tests with $c = 1, 2, 3$ in the specific example in which $\mu_0 = 4$, $n = 15$, and $\sigma^2 = 9$. ◀

Since the possibility of error exists in virtually every testing problem, we should consider what kinds of errors we might make. For each value of $\theta \in \Omega_0$, the decision

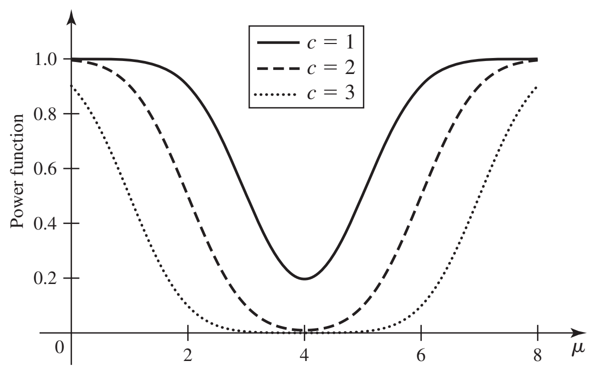


Figure 9.1 Power functions of three different tests in Example 9.1.5.

to reject H_0 is an incorrect decision. Similarly, for each value of $\theta \in \Omega_1$, the decision not to reject H_0 is an incorrect decision.

Definition 9.1.7 Type I/II Error. An erroneous decision to reject a true null hypothesis is a *type I error*, or an error of the first kind. An erroneous decision not to reject a false null hypothesis is called a *type II error*, or an error of the second kind.

In terms of the power function, if $\theta \in \Omega_0$, $\pi(\theta|\delta)$ is the probability that the statistician will make a type I error. Similarly, if $\theta \in \Omega_1$, $1 - \pi(\theta|\delta)$ is the probability of making a type II error. Of course, either $\theta \in \Omega_0$ or $\theta \in \Omega_1$, but not both. Hence, only one type of error is possible conditional on θ, but we never know which it is.

If we have our choice between several tests, we would like to choose a test δ that has small probability of error. That is, we would like the power function $\pi(\theta|\delta)$ to be low for values of $\theta \in \Omega_0$, and we would like $\pi(\theta|\delta)$ to be high for $\theta \in \Omega_1$. Generally, these two goals work against each other. That is, if we choose δ to make $\pi(\theta|\delta)$ small for $\theta \in \Omega_0$, we will usually find that $\pi(\theta|\delta)$ is small for $\theta \in \Omega_1$ as well. For example, the test procedure δ_0 that never rejects H_0, regardless of what data are observed, will have $\pi(\theta|\delta_0) = 0$ for all $\theta \in \Omega_0$. However, for this procedure $\pi(\theta|\delta_0) = 0$ for all $\theta \in \Omega_1$ as well. Similarly, the test δ_1 that always rejects H_0 will have $\pi(\theta|\delta_1) = 1$ for all $\theta \in \Omega_1$, but it will also have $\pi(\theta|\delta_1) = 1$ for all $\theta \in \Omega_0$. Hence, there is a need to strike an appropriate balance between the two goals of low power in Ω_0 and high power in Ω_1.

The most popular method for striking a balance between the two goals is to choose a number α_0 between 0 and 1 and require that

$$\pi(\theta|\delta) \le \alpha_0, \quad \text{for all } \theta \in \Omega_0. \tag{9.1.6}$$

Then, among all tests that satisfy (9.1.6), the statistician seeks a test whose power function is as high as can be obtained for $\theta \in \Omega_1$. This method is discussed in Sections 9.2 and 9.3. Another method of balancing the probabilities of type I and type II errors is to minimize a linear combination of the different probabilities of error. We shall discuss this method in Sec. 9.2 and again in Sec. 9.8.

Note: Choosing Null and Alternative Hypotheses. If one chooses to balance type I and type II error probabilities by requiring (9.1.6), then one has introduced an asymmetry in the treatment of the null and alternative hypotheses. In most testing problems, such asymmetry can be quite natural. Generally, one of the two errors (type I or type II) is more costly or less palatable in some sense. It would make sense to put tighter controls on the probability of the more serious error. For this reason, one generally arranges the null and alternative hypotheses so that type I error is the error most to be avoided. For cases in which neither hypothesis is naturally the null, switching the names of null and alternative hypotheses can have a variety of different effects on the results of testing procedures. (See Exercise 21 in this section.)

Example 9.1.6 Egyptian Skulls. In Example 9.1.2, suppose that the experimenters have a theory saying that skull breadths should increase (albeit slightly) over long periods of time. If μ is the mean breadth of skulls from 4000 B.C. and 140 is the mean breadth of modern-day skulls, the theory would say $\mu < 140$. The experimenters could mistakenly claim that the data support their theory ($\mu < 140$) when, in fact, $\mu > 140$, or they might mistakenly claim that the data fail to support their theory ($\mu > 140$) when, in fact, $\mu < 140$. In scientific studies, it is common to treat the false confirmation of one's own theory as a more serious error than falsely failing to confirm ones' own theory. This would mean type I error should be to say that $\mu < 140$ (confirm the theory, reject H_0) when, in fact, $\mu > 140$ (theory is false, H_0 is true). Traditionally, one includes the

endpoints of interval hypotheses in the null, so we would formulate the hypotheses to be tested as

$$H_0:\quad \mu \geq 140,$$
$$H_1:\quad \mu < 140,$$

as we did in Example 9.1.2. ◀

The quantities in Eq. (9.1.6) play a fundamental role in hypothesis testing and have special names.

Definition 9.1.8 Level/Size. A test that satisfies (9.1.6) is called a *level* α_0 *test*, and we say that the test has *level of significance* α_0. In addition, the *size* $\alpha(\delta)$ of a test δ is defined as follows:

$$\alpha(\delta) = \sup_{\theta \in \Omega_0} \pi(\theta|\delta). \tag{9.1.7}$$

The following results are immediate consequences of Definition 9.1.8.

Corollary 9.1.1 A test δ is a level α_0 test if and only if its size is at most α_0 (i.e., $\alpha(\delta) \leq \alpha_0$). If the null hypothesis is simple, that is, $H_0 : \theta = \theta_0$, then the size of δ will be $\alpha(\delta) = \pi(\theta_0|\delta)$. ■

Example 9.1.7 Testing Hypotheses about a Uniform Distribution. Suppose that a random sample $X_1, \ldots, X_n$ is taken from the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown ($\theta > 0$); and suppose also that it is desired to test the following hypotheses:

$$\begin{aligned} &H_0: 3 \leq \theta \leq 4,\\ &H_1: \theta < 3 \text{ or } \theta > 4. \end{aligned} \tag{9.1.8}$$

We know from Example 6.5.15 that the M.L.E. of θ is $Y_n = \max\{X_1, \ldots, X_n\}$. Although Y_n must be less than θ, there is a high probability that Y_n will be close to θ if the sample size n is fairly large. For illustrative purposes, suppose that the test δ does not reject H_0 if $2.9 < Y_n < 4$, and δ rejects H_0 if Y_n does not lie in this interval. Thus, the critical region of the test δ contains all the values of $X_1, \ldots, X_n$ for which either $Y_n \leq 2.9$ or $Y_n \geq 4$. In terms of the test statistic Y_n, the rejection region is the union of two intervals $(-\infty, 2.9] \cup [4, \infty)$.

The power function of δ is specified by the relation

$$\pi(\theta|\delta) = \Pr(Y_n \leq 2.9|\theta) + \Pr(Y_n \geq 4|\theta).$$

If $\theta \leq 2.9$, then $\Pr(Y_n \leq 2.9|\theta) = 1$ and $\Pr(Y_n \geq 4|\theta) = 0$. Therefore, $\pi(\theta|\delta) = 1$ if $\theta \leq 2.9$. If $2.9 < \theta \leq 4$, then $\Pr(Y_n \leq 2.9|\theta) = (2.9/\theta)^n$ and $\Pr(Y_n \geq 4|\theta) = 0$. In this case, $\pi(\theta|\delta) = (2.9/\theta)^n$. Finally, if $\theta > 4$, then $\Pr(Y_n \leq 2.9|\theta) = (2.9/\theta)^n$ and $\Pr(Y_n \geq 4|\theta) = 1 - (4/\theta)^n$. In this case, $\pi(\theta|\delta) = (2.9/\theta)^n + 1 - (4/\theta)^n$. The power function $\pi(\theta|\delta)$ is sketched in Fig. 9.2.

By Eq. (9.1.7), the size of δ is $\alpha(\delta) = \sup_{3 \leq \theta \leq 4} \pi(\theta|\delta)$. It can be seen from Fig. 9.2 and the calculations just given that $\alpha(\delta) = \pi(3|\delta) = (29/30)^n$. In particular, if the sample size is $n = 68$, then the size of δ is $(29/30)^{68} = 0.0997$. So δ is a level α_0 test for every level of significance $\alpha_0 \geq 0.0997$. ◀

Making a Test Have a Specific Significance Level

Suppose that we wish to test the hypotheses

$$H_0:\quad \theta \in \Omega_0,$$
$$H_1:\quad \theta \in \Omega_1.$$

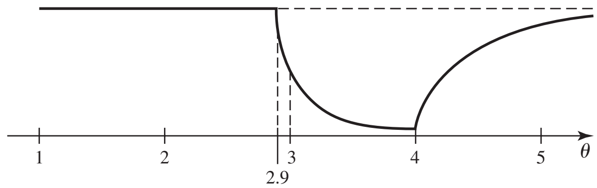


Figure 9.2 The power function $\pi(\theta|\delta)$ in Example 9.1.7.

Let T be a test statistic, and suppose that our test will reject the null hypothesis if $T \geq c$, for some constant c. Suppose also that we desire our test to have the level of significance α_0. The power function of our test is $\pi(\theta|\delta) = \Pr(T \geq c|\theta)$, and we want

$$\sup_{\theta \in \Omega_0} \Pr(T \geq c|\theta) \leq \alpha_0. \tag{9.1.9}$$

It is clear that the power function, and hence the left side of (9.1.9), are nonincreasing functions of c. Hence, (9.1.9) will be satisfied for large values of c, but not for small values. If we want the power function to be as large as possible for $\theta \in \Omega_1$, we should make c as small as we can while still satisfying (9.1.9). If T has a continuous distribution, then it is usually simple to find an appropriate c.

Example 9.1.8 Testing Hypotheses about the Mean of a Normal Distribution with Known Variance. In Example 9.1.5, our test is to reject $H_0: \mu = \mu_0$ if $|\overline{X}_n - \mu_0| \geq c$. Since the null hypothesis is simple, the left side of (9.1.9) reduces to the probability (assuming that $\mu = \mu_0$) that $|\overline{X}_n - \mu_0| \geq c$. Since $Y = \overline{X}_n - \mu_0$ has the normal distribution with mean 0 and variance σ^2/n when $\mu = \mu_0$, we can find a value c that makes the size exactly α_0 for each α_0. Figure 9.3 shows the p.d.f. of Y and the size of the test indicated as the shaded area under the p.d.f. Since the normal p.d.f. is symmetric around the mean (0 in this case), the two shaded areas must be the same, namely, $\alpha_0/2$. This means that c must be the $1 - \alpha_0/2$ quantile of the distribution of Y. This quantile is $c = \Phi^{-1}(1 - \alpha_0/2)\sigma n^{-1/2}$.

When testing hypotheses about the mean of a normal distribution, it is traditional to rewrite this test in terms of the statistic

$$Z = n^{1/2}\frac{\overline{X}_n - \mu_0}{\sigma}. \tag{9.1.10}$$

Then the test rejects H_0 if $|Z| \geq \Phi^{-1}(1 - \alpha_0/2)$. ◀

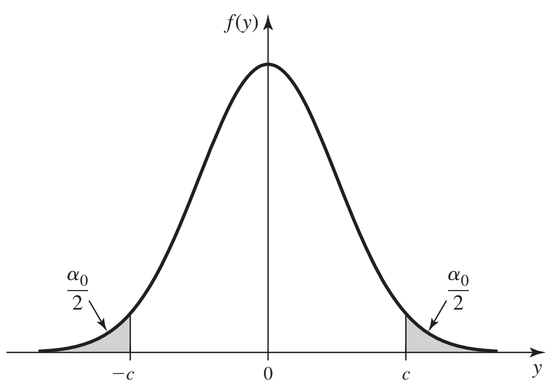


Figure 9.3 The p.d.f. of $Y = \overline{X}_n - \mu_0$ given $\mu = \mu_0$ for Example 9.1.8. The shaded areas represent the probability that $|Y| \geq c$.

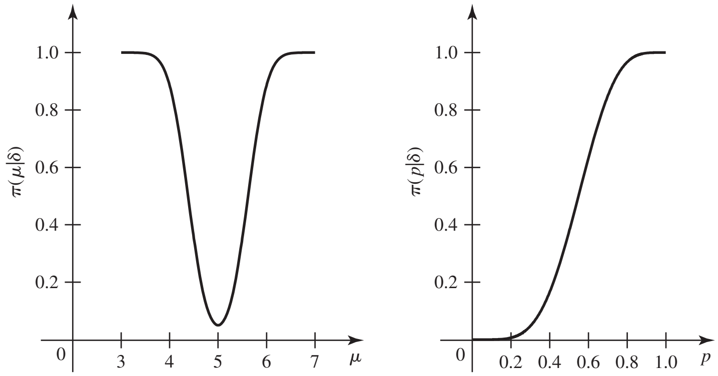


Figure 9.4 Power functions of two tests. The plot on the left is the power function of the test from Example 9.1.8 with $n = 10$, $\mu_0 = 5$, $\sigma = 1$, and $\alpha_0 = 0.05$. The plot on the right is the power function of the test from Example 9.1.9 with $n = 10$, $p_0 = 0.3$, and $\alpha_0 = 0.1$.

Example 9.1.9

Testing Hypotheses about a Bernoulli Parameter. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with parameter p. Suppose that we wish to test the hypotheses

$$\begin{aligned} H_0&: \quad p \le p_0, \\ H_1&: \quad p > p_0. \end{aligned} \tag{9.1.11}$$

Let $Y = \sum_{i=1}^{n} X_i$, which has the binomial distribution with parameters n and p. The larger p is, the larger we expect Y to be. So, suppose that we choose to reject H_0 if $Y \ge c$, for some constant c. Suppose also that we want the size of the test to be as close to α_0 as possible without exceeding α_0. It is easy to check that $\Pr(Y \ge c|p)$ is an increasing function of p; hence, the size of the test will be $\Pr(Y \ge c|p = p_0)$. So, c should be the smallest number such that $\Pr(Y \ge c|p = p_0) \le \alpha_0$. For example, if $n = 10$, $p_0 = 0.3$, and $\alpha_0 = 0.1$, we can use the table of binomial probabilities in the back of this book to determine c. We can compute $\sum_{y=6}^{10} \Pr(Y = y|p = 0.3) = 0.0473$ and $\sum_{y=5}^{10} \Pr(Y = y|p = 0.3) = 0.1503$. In order to keep the size of the test at most 0.1, we must choose $c > 5$. Every value of c in the interval (5, 6] produces the same test, since Y takes only integer values. ◄

Whenever we choose a test procedure, we should also examine the power function. If one has made a good choice, then the power function should generally be larger for $\theta \in \Omega_1$ than for $\theta \in \Omega_0$. Also, the power function should increase as θ moves away from Ω_0. For example, Fig. 9.4 shows plots of the power functions for two of the examples in this section. In both cases, the power function increases as the parameter moves away from Ω_0.

The *p*-value

Example 9.1.10

Testing Hypotheses about the Mean of a Normal Distribution with Known Variance. In Example 9.1.8, suppose that we choose to test the null hypothesis at level $\alpha_0 = 0.05$. We would then compute the test statistic in Eq. (9.1.10) and reject H_0 if $Z \ge \Phi^{-1}(1 - 0.05/2) = 1.96$. For example, suppose that $Z = 2.78$ is observed. Then we would reject H_0. Suppose that we were to report the result by saying that we rejected H_0 at level 0.05. What would another statistician, who felt it more appropriate to test the null hypothesis at a different level, be able to do with this report? ◄

The result of a test of hypotheses might appear to be a rather inefficient use of our data. For instance, in Example 9.1.10, we decided to reject H_0 at level $\alpha_0 = 0.05$ if the statistic Z in Eq. (9.1.10) is at least 1.96. This means that whether we observe $Z = 1.97$ or $Z = 6.97$, we shall report the same result, namely, that we rejected H_0 at level 0.05. The report of the test result does not carry any sense of how close we were to making the other decision. Furthermore, if another statistician chooses to use a size 0.01 test, then she would not reject H_0 with $Z = 1.97$, but she would reject H_0 with $Z = 6.97$. What would she do with $Z = 2.78$?

For these reasons, an experimenter does not typically choose a value of α_0 in advance of the experiment and then simply report whether or not H_0 was rejected at level α_0. In many fields of application, it has become standard practice to report, in addition to the observed value of the appropriate test statistic such as Z, *all* the values of α_0 for which the level α_0 test would lead to the rejection of H_0.

Example 9.1.11 **Testing Hypotheses about the Mean of a Normal Distribution with Known Variance.** As the observed value of Z in Example 9.1.8 is 2.78, the hypothesis H_0 would be rejected for every level of significance α_0 such that $2.78 \geq \Phi^{-1}(1 - \alpha_0/2)$. Using the table of the normal distribution given at the end of this book, this inequality translates to $\alpha_0 \geq 0.0054$. The value 0.0054 is called the *p-value* for the observed data and the tested hypotheses. Since $0.01 > 0.0054$, the statistician who wanted to test the hypotheses at level 0.01 would also reject H_0. ◀

Definition 9.1.9 ***p*-value.** In general, the *p-value* is the smallest level α_0 such that we would reject the null-hypothesis at level α_0 with the observed data.

An experimenter who rejects a null hypothesis if and only if the p-value is at most α_0 is using a test with level of significance α_0. Similarly, an experimenter who wants a level α_0 test will reject the null hypothesis if and only if the p-value is at most α_0. For this reason, the p-value is sometimes called the *observed level of significance*.

An experimenter in Example 9.1.10 would typically report that the observed value of Z was 2.78 and that the corresponding p-value was 0.0054. It is then said that the observed value of Z is *just significant* at the level of significance 0.0054. One advantage to the experimenter of reporting experimental results in this manner is that he does not need to select beforehand an arbitrary level of significance α_0 at which to carry out the test. Also, when a reader of the experimenter's report learns that the observed value of Z was just significant at the level of significance 0.0054, she immediately knows that H_0 would be rejected for every larger value of α_0 and would not be rejected for any smaller value.

Calculating p-values If all of our tests are of the form "reject the null hypothesis when $T \geq c$" for a single test statistic T, there is a straightforward way to compute p-values. For each t, let δ_t be the test that rejects H_0 if $T \geq t$. Then the p-value when $T = t$ is observed is the size of the test δ_t. (See Exercise 18.) That is, the p-value equals

$$\sup_{\theta \in \Omega_0} \pi(\theta|\delta_t) = \sup_{\theta \in \Omega_0} \Pr(T \geq t|\theta). \tag{9.1.12}$$

Typically, $\pi(\theta|\delta_t)$ is maximized at some θ_0 on the boundary between Ω_0 and Ω_1. Because the p-value is calculated as a probability in the upper tail of the distribution of T, it is sometimes called a *tail area*.

Example 9.1.12 **Testing Hypotheses about a Bernoulli Parameter.** For testing the hypotheses (9.1.11) in Example 9.1.9, we used a test that rejects H_0 if $Y \geq c$. The p-value, when $Y = y$ is observed, will be $\sup_{p \leq p_0} \Pr(Y \geq y|p)$. In this example, it is easy to see that $\Pr(Y \geq y|p)$ increases as a function of p. Hence, the p-value is $\Pr(Y \geq y|p = p_0)$. For example, let $p_0 = 0.3$ and $n = 10$. If $Y = 6$ is observed, then $\Pr(Y \geq 6|p = 0.3) = 0.0473$, as we calculated in Example 9.1.9. ◄

The calculation of the p-value is more complicated when the test cannot be put into the form "reject H_0 if $T \geq c$." In this text, we shall calculate p-values only for tests that do have this form.

Equivalence of Tests and Confidence Sets

Example 9.1.13 **Rain from Seeded Clouds.** In Examples 8.5.5 and 8.5.6, we found a coefficient γ one-sided (lower limit) confidence interval for μ, the mean log-rainfall from seeded clouds. For $\gamma = 0.9$, the observed interval is $(4.727, \infty)$. One of the controversial interpretations of this interval is that we have confidence 0.9 (whatever that means) that $\mu > 4.727$. Although this statement is deliberately ambiguous and difficult to interpret, it sounds as if it could help us address the problem of testing the hypotheses $H_0 : \mu \leq 4$ versus $H_1 : \mu > 4$. Does the fact that 4 is not in the observed coefficient 0.9 confidence interval tell us anything about whether or not we should reject H_0 at some significance level or other? ◄

We shall now illustrate how confidence intervals (see Sec. 8.5) can be used as an alternative method to report the results of a test of hypotheses. In particular, we shall show that a coefficient γ confidence set (a generalization of confidence interval to be defined shortly) can be thought of as a set of null hypotheses that would not be rejected at significance level $1 - \gamma$.

Theorem 9.1.1 **Defining Confidence Sets from Tests.** Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter θ. Let $g(\theta)$ be a function, and suppose that for each possible value g_0 of $g(\theta)$, there is a level α_0 test δ_{g_0} of the hypotheses

$$H_{0,g_0} : g(\theta) = g_0, \quad H_{1,g_0} : g(\theta) \neq g_0. \tag{9.1.13}$$

For each possible value $\boldsymbol{x}$ of $\boldsymbol{X}$, define

$$\omega(\boldsymbol{x}) = \{g_0 : \delta_{g_0} \text{ does not reject } H_{0,g_0} \text{ if } \boldsymbol{X} = \boldsymbol{x} \text{ is observed}\}. \tag{9.1.14}$$

Let $\gamma = 1 - \alpha_0$. Then, the random set $\omega(\boldsymbol{X})$ satisfies

$$\Pr[g(\theta_0) \in \omega(\boldsymbol{X})|\theta = \theta_0] \geq \gamma. \tag{9.1.15}$$

for all $\theta_0 \in \Omega$.

Proof Let θ_0 be an arbitrary element of Ω, and define $g_0 = g(\theta_0)$. Because δ_{g_0} is a level α_0 test, we know that

$$\Pr[\delta_{g_0} \text{ does not reject } H_{0,g_0}|\theta = \theta_0] \geq 1 - \alpha_0 = \gamma. \tag{9.1.16}$$

For each $\boldsymbol{x}$, we know that $g(\theta_0) \in \omega(\boldsymbol{x})$ if and only if the test δ_{g_0} does not reject H_{0,g_0} when $\boldsymbol{X} = \boldsymbol{x}$ is observed. It follows that the left-hand side of Eq. (9.1.15) is the same as the left-hand side of Eq. (9.1.16). ■

Definition 9.1.10 Confidence Set. If a random set $\omega(\boldsymbol{X})$ satisfies (9.1.15) for every $\theta_0 \in \Omega$, we call it a *coefficient* γ *confidence set for* $g(\theta)$. If the inequality in (9.1.15) is equality for all θ_0, then we call the confidence set *exact*.

A confidence set is a generalization of the concept of a confidence interval introduced in Sec. 8.5. What Theorem 9.1.1 shows is that a collection of level α_0 tests of the hypotheses (9.1.13) can be used to construct a coefficient $\gamma = 1 - \alpha_0$ confidence set for $g(\theta)$. The reverse construction is also possible.

Theorem 9.1.2 Defining Tests from Confidence Sets. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter θ. Let $g(\theta)$ be a function of θ, and let $\omega(\boldsymbol{X})$ be a coefficient γ confidence set for $g(\theta)$. For each possible value g_0 of $g(\theta)$, construct the following test δ_{g_0} of the hypotheses in Eq. (9.1.13): δ_{g_0} does not reject H_{0,g_0} if and only if $g_0 \in \omega(\boldsymbol{X})$. Then δ_{g_0} is a level $\alpha_0 = 1 - \gamma$ test of the hypotheses in Eq. (9.1.13).

Proof Because $\omega(\boldsymbol{X})$ is a coefficient γ confidence set for $g(\theta)$, it satisfies Eq. (9.1.15) for all $\theta_0 \in \Omega$. As in the proof of Theorem 9.1.1, the left-hand sides of Eqs. (9.1.15) and (9.1.16) are the same, which makes δ_{g_0} a level α_0 test. ■

Example 9.1.14 A Confidence Interval for the Mean of a Normal Distribution. Consider the test found in Example 9.1.8 for the hypotheses (9.1.2). Let $\alpha_0 = 1 - \gamma$. The size α_0 test δ_{μ_0} is to reject H_0 if $|\overline{X}_n - \mu_0| \ge \Phi^{-1}(1 - \alpha_0/2)\sigma n^{-1/2}$. If $\overline{X}_n = \overline{x}_n$ is observed, the set of μ_0 such that we would not reject H_0 is the set of μ_0 such that

$$|\overline{x}_n - \mu_0| < \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)\sigma n^{-1/2}.$$

This inequality easily translates to

$$\overline{x}_n - \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)\sigma n^{-1/2} < \mu_0 < \overline{x}_n + \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)\sigma n^{-1/2}.$$

The coefficient γ confidence interval becomes

$$(A, B) = \left(\overline{X}_n - \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)\sigma n^{-1/2}, \overline{X}_n + \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)\sigma n^{-1/2}\right).$$

It is easy to check that $\Pr(A < \mu_0 < B|\mu = \mu_0) = \gamma$ for all μ_0. This confidence interval is exact. ◄

Example 9.1.15 Constructing a Test from a Confidence Interval. In Sec. 8.5, we learned how to construct a confidence interval for the unknown mean of a normal distribution when the variance was also unknown. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. In this case, the parameter is $\theta = (\mu, \sigma^2)$, and we are interested in $g(\theta) = \mu$. In Sec. 8.5, we used the statistics

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \sigma' = \left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right)^{1/2}. \tag{9.1.17}$$

The coefficient γ confidence interval for $g(\theta)$ is the interval

$$\left(\overline{X}_n - T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma'}{n^{1/2}}, \overline{X}_n + T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma'}{n^{1/2}}\right), \tag{9.1.18}$$

where $T_{n-1}^{-1}(\cdot)$ is the quantile function of the t distribution with $n-1$ degrees of freedom. For each μ_0, we can use this interval to find a level $\alpha_0 = 1-\gamma$ test of the hypotheses

$$H_0: \quad \mu = \mu_0,$$
$$H_1: \quad \mu \neq \mu_0.$$

The test will reject H_0 if μ_0 is not in the interval (9.1.18). A little algebra shows that μ_0 is not in the interval (9.1.18) if and only if

$$\left| n^{1/2}\frac{\overline{X}_n - \mu_0}{\sigma'} \right| \geq T^{-1}\left(\frac{1+\gamma}{2}\right).$$

This test is identical to the t test that we shall study in more detail in Sec. 9.5. ◀

One-Sided Confidence Intervals and Tests Theorems 9.1.1 and 9.1.2 establish the equivalence between confidence sets and tests of hypotheses of the form (9.1.13). It is often necessary to test other forms of hypotheses, and it would be nice to have versions of Theorems 9.1.1 and 9.1.2 to deal with these cases. Example 9.1.13 is one such case in which the hypotheses are of the form

$$H_{0,g_0}: g(\theta) \leq g_0, \qquad H_{1,g_0}: g(\theta) > g_0. \tag{9.1.19}$$

Theorem 9.1.1 extends immediately to such cases. We leave the proof of Theorem 9.1.3 to the reader.

Theorem 9.1.3 One-Sided Confidence Intervals from One-Sided Tests. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter θ. Let $g(\theta)$ be a real-valued function, and suppose that for each possible value g_0 of $g(\theta)$, there is a level α_0 test δ_{g_0} of the hypotheses (9.1.19). For each possible value $\boldsymbol{x}$ of $\boldsymbol{X}$, define $\omega(\boldsymbol{x})$ by Eq. (9.1.14). Let $\gamma = 1 - \alpha_0$. Then the random set $\omega(\boldsymbol{X})$ satisfies Eq. (9.1.15) for all $\theta_0 \in \Omega$. ■

Example 9.1.16 One-Sided Confidence Interval for a Bernoulli Parameter. In Example 9.1.9, we showed how to construct a level α_0 test of the one-sided hypotheses (9.1.11). Let $Y = \sum_{i=1}^{n} X_i$. The test rejects H_0 if $Y \geq c(p_0)$ where $c(p_0)$ is the smallest number c such that $\Pr(Y \geq c | p = p_0) \leq \alpha_0$. After observing the data $\boldsymbol{X}$, we can check, for each p_0, whether or not we reject H_0. That is, for each p_0 we check whether or not $Y \geq c(p_0)$. All those p_0 for which $Y < c(p_0)$ (i.e., we don't reject H_0) will form an interval $\omega(\boldsymbol{X})$. This interval will satisfy $\Pr(p_0 \in \omega(\boldsymbol{X}) | p = p_0) \geq 1 - \alpha_0$ for all p_0. For example, suppose that $n = 10$, $\alpha_0 = 0.1$, and $Y = 6$ is observed. In order not to reject $H_0: p \leq p_0$ at level 0.1, we must have a rejection region that does not contain 6. This will happen if and only if $\Pr(Y \geq 6 | p = p_0) > 0.1$. By trying various values of p_0, we find that this inequality holds for all $p_0 > 0.3542$. So, if $Y = 6$ is observed, our coefficient 0.9 confidence interval is (0.3542, 1). Notice that 0.3 is not in the interval, so we would reject $H_0: p \leq 0.3$ with a level 0.1 test as we did in Example 9.1.9. For other observed values $Y = y$, the confidence intervals will all be of the form $(q(y), 1)$ where $q(y)$ can be computed as outlined in Exercise 17. For $n = 10$ and $\alpha_0 = 0.1$, the values of $q(y)$ are

y	0	1	2	3	4	5	6	7	8	9	10
$q(y)$	0	0.0104	0.0545	0.1158	0.1875	0.2673	0.3542	0.4482	0.5503	0.6631	0.7943

This confidence interval is not exact. ◀

Unfortunately, Theorem 9.1.2 does not immediately extend to one-sided hypotheses for the following reason. The size of a one-sided test for hypotheses of the form (9.1.19) depends on *all* of the values of θ such that $g(\theta) \le g_0$, not just on those for which $g(\theta) = g_0$. In particular, the size of the test δ_{g_0} defined in Theorem 9.1.2 is

$$\sup_{\{\theta: g(\theta) \le g_0\}} \Pr[g_0 \notin \omega(\boldsymbol{X})|\theta]. \tag{9.1.20}$$

The confidence coefficient, on the other hand, is

$$1 - \sup_{\{\theta: g(\theta) = g_0\}} \Pr[g_0 \notin \omega(\boldsymbol{X})|\theta].$$

If we could prove that the supremum in Eq. (9.1.20) occurred at a θ for which $g(\theta) = g_0$, then the size of the test would be 1 minus the confidence coefficient. Most of the cases with which we shall deal in this book will have the property that the supremum in Eq. (9.1.20) does indeed occur at a θ for which $g(\theta) = g_0$. Example 9.1.16 is one such case. Example 9.1.13 is another. The following example is the general version of what we need in Example 9.1.13.

Example 9.1.17 One-Sided Tests and Confidence Intervals for a Normal Mean with Unknown Variance. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. Here $\theta = (\mu, \sigma^2)$. Let $g(\theta) = \mu$. In Theorem 8.5.1, we found that

$$\left(\overline{X}_n - T_{n-1}^{-1}(\gamma)\,\frac{\sigma'}{n^{1/2}}, \infty\right) \tag{9.1.21}$$

is a one-sided coefficient γ confidence interval for $g(\theta)$. Now, suppose that we use this interval to test hypotheses. We shall reject the null hypothesis that $\mu = \mu_0$ if μ_0 is not in the interval (9.1.21). It is easy to see that μ_0 is not in the interval (9.1.21) if and only if $\overline{X}_n \ge \mu_0 + \sigma' n^{-1/2} T_{n-1}^{-1}(\gamma)$. Such a test would seem to make sense for testing the hypotheses

$$H_0\colon \mu \le \mu_0, \qquad H_1\colon \mu > \mu_0. \tag{9.1.22}$$

In particular, in Example 9.1.13, the fact that 4 is not in the observed confidence interval means that the test constructed above (with $\mu_0 = 4$ and $\gamma = 0.9$) would reject $H_0: \mu \le 4$ at level $\alpha_0 = 0.1$. ◀

The test constructed in Example 9.1.17 is another t test that we shall study in Sec. 9.5. In particular, we will show in Sec. 9.5 that this t test is a level $1 - \gamma$ test. In Exercise 19, you can find the one-sided confidence interval that corresponds to testing the reverse hypotheses.

Likelihood Ratio Tests

A very popular form of hypothesis test is the likelihood ratio test. We shall give a partial theoretical justification for likelihood ratio tests in Sec. 9.2. Such tests are based on the likelihood function $f_n(\boldsymbol{x}|\theta)$. (See Definition 7.2.3 on page 390.) The likelihood function tends to be highest near the true value of θ. Indeed, this is why maximum likelihood estimation works well in so many cases. Now, suppose that we wish to test the hypotheses

$$\begin{aligned} H_0&\colon \quad \theta \in \Omega_0, \\ H_1&\colon \quad \theta \in \Omega_1. \end{aligned} \tag{9.1.23}$$

In order to compare these two hypotheses, we might wish to see whether the likelihood function is higher on Ω_0 or on Ω_1, and if not, how much smaller the likelihood

function is on Ω_0. When we computed M.L.E.'s, we maximized the likelihood function over the entire parameter space Ω. In particular, we calculated $\sup_{\theta\in\Omega} f_n(\boldsymbol{x}|\theta)$. If we restrict attention to H_0, then we can compute the largest value of the likelihood among those parameter values in Ω_0: $\sup_{\theta\in\Omega_0} f_n(\boldsymbol{x}|\theta)$. The ratio of these two suprema can then be used for testing the hypotheses (9.1.23).

Definition 9.1.11 Likelihood Ratio Test. The statistic

$$\Lambda(\boldsymbol{x}) = \frac{\sup_{\theta\in\Omega_0} f_n(\boldsymbol{x}|\theta)}{\sup_{\theta\in\Omega} f_n(\boldsymbol{x}|\theta)} \tag{9.1.24}$$

is called the *likelihood ratio statistic*. A *likelihood ratio test* of hypotheses (9.1.23) is to reject H_0 if $\Lambda(\boldsymbol{x}) \le k$ for some constant k.

In words, a likelihood ratio test rejects H_0 if the likelihood function on Ω_0 is sufficiently small compared to the likelihood function on all of Ω. Generally, k is chosen so that the test has a desired level α_0, if that is possible.

Example 9.1.18 Likelihood Ratio Test of Two-Sided Hypotheses about a Bernoulli Parameter. Suppose that we shall observe Y, the number of successes in n independent Bernoulli trials with unknown parameter θ. Consider the hypotheses $H_0 : \theta = \theta_0$ versus $H_0 : \theta \ne \theta_0$. After the value $Y = y$ has been observed, the likelihood function is

$$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

In this case, $\Omega_0 = \{\theta_0\}$ and $\Omega = [0, 1]$. The likelihood ratio statistic is

$$\Lambda(y) = \frac{\theta_0^y(1-\theta_0)^{n-y}}{\sup_{\theta\in[0,1]}\theta^y(1-\theta)^{n-y}}. \tag{9.1.25}$$

The supremum in the denominator of Eq. (9.1.25) can be found as in Example 7.5.4. The maximum occurs where θ equals the M.L.E., $\hat{\theta} = y/n$. So,

$$\Lambda(y) = \left(\frac{n\theta_0}{y}\right)^y \left(\frac{n(1-\theta_0)}{n-y}\right)^{n-y}.$$

It is not difficult to see that $\Lambda(y)$ is small for y near 0 and near n and largest near $y = n\theta_0$. As a specific example, suppose that $n = 10$ and $\theta_0 = 0.3$. Table 9.1 shows the 11 possible values of $\Lambda(y)$ for $y = 0, \ldots, 10$. If we desired a test with level of significance α_0, we would order the values of y according to values of $\Lambda(y)$ from smallest to largest and choose k so that the sum of the probabilities $\Pr(Y = y|\theta = 0.3)$ corresponding to those values of y with $\Lambda(y) \le k$ was at most α_0. For example, if $\alpha_0 = 0.05$, we see from Table 9.1 that we can add up the probabilities corresponding to $y = 10, 9, 8, 7, 0$ to get 0.039. But if we include $y = 6$, corresponding to the next smallest value of $\Lambda(y)$, the sum jumps to 0.076, which is too large. The set of $y \in \{10, 9, 8, 7, 0\}$ corresponds to $\Lambda(y) \le k$ for every k in the half-open interval [0.028, 0.147). The size of the test that rejects H_0 when $y \in \{10, 9, 8, 7, 0\}$ is 0.039. ◂

Likelihood Ratio Tests with Large Samples

Likelihood ratio tests are most popular in problems involving large sample sizes. The following result, whose precise statement and proof are beyond the scope of this text, shows how to use them in such cases.

Table 9.1 Values of the likelihood ratio statistic in Example 9.1.18

y	0	1	2	3	4	5	6	7	8	9	10
$\Lambda(y)$	0.028	0.312	0.773	1.000	0.797	0.418	0.147	0.034	0.005	3×10^{-4}	6×10^{-6}
$\Pr(Y = y\|\theta = 0.3)$	0.028	0.121	0.233	0.267	0.200	0.103	0.037	0.009	0.001	1×10^{-4}	6×10^{-6}

Theorem 9.1.4 Large-Sample Likelihood Ratio Tests. Let Ω be an open subset of p-dimensional space, and suppose that H_0 specifies that k coordinates of θ are equal to k specific values. Assume that H_0 is true and that the likelihood function satisfies the conditions needed to prove that the M.L.E. is asymptotically normal and asymptotically efficient. (See page 523.) Then, as $n \to \infty$, $-2 \log \Lambda(\boldsymbol{X})$ converges in distribution to the χ^2 distribution with k degrees of freedom. ■

Example 9.1.19 Likelihood Ratio Test of Two-Sided Hypotheses about a Bernoulli Parameter. We shall apply the idea in Theorem 9.1.4 to the case at the end of Example 9.1.18. Set $\Omega = (0, 1)$ so that $p = 1$ and $k = 1$. To get an approximate level α_0 test, we would reject H_0 if $-2 \log \Lambda(y)$ is greater than the $1 - \alpha_0$ quantile of the χ^2 distribution with one degree of freedom. With $\alpha_0 = 0.05$, this quantile is 3.841. By taking logarithms of the numbers in the $\Lambda(y)$ row of Table 9.1, one sees that $-2 \log \Lambda(y) > 3.841$ for $y \in \{10, 9, 8, 7, 0\}$. Rejecting H_0 when $-2 \log \Lambda(y) > 3.841$ is then the same test as we constructed in Example 9.1.18. ◄

Theorem 9.1.4 can also be applied if the null hypothesis specifies that a collection of k functions of θ are equal to k specific values. For example, suppose that the parameter is $\theta = (\mu, \sigma^2)$, and we wish to test $H_0 : (\mu - 2)/\sigma = 1$ versus $H_1 : (\mu - 2)/\sigma \neq 1$. We could first transform to the equivalent parameter $\theta' = ([\mu - 2]/\sigma, \sigma)$ and then apply Theorem 9.1.4. Because of the invariance property of M.L.E.'s (Theorem 7.6.1, which extends to multidimensional parameters) one does not actually need to perform the transformation in order to compute Λ. One merely needs to maximize the likelihood function over the two sets Ω_0 and Ω and take the ratio.

On a final note, one must be careful not to apply Theorem 9.1.4 to problems of one-sided hypothesis testing. In such cases, the $\Lambda(\boldsymbol{X})$ usually has a distribution that is neither discrete nor continuous and doesn't converge to a χ^2 distribution. Also, Theorem 9.1.4 fails to apply when the parameter space Ω is a closed set and the null hypothesis is that θ takes a value on the boudary of Ω.

Hypothesis-Testing Terminology

We noted after Definition 9.1.1 that there is asymmetry in the terminology with regard to choosing between hypotheses. Both choices are stated relative to H_0, namely, to reject H_0 or not to reject H_0. When hypothesis testing was first being developed, there was controversy over whether alternative hypotheses should even be formulated. Focus centered on null hypotheses and whether or not to reject them. The operational meaning of "do not reject H_0" has never been articulated clearly. In particular, it does not mean that we should accept H_0 as true in any sense. Nor does it mean that we are necessarily more confident that H_0 is true than that it is false. For

that matter, "reject H_0" does not mean that we are more confident that H_0 is false than that it is true.

Part of the problem is that hypothesis testing is set up as if it were a statistical decision problem, but neither a loss function nor a utility function is involved. Hence, we are not weighing the relative likelihoods of various hypotheses against the costs or benefits of making various decisions. In Sec. 9.8, we shall illustrate one method for treating the hypothesis-testing problem as a statistical decision problem. Many, but not all, of the popular testing procedures will turn out to have interpretations in the framework of decision problems. In the remainder of this chapter, we shall continue to develop the theory of hypothesis testing as it is generally practiced.

There are two other points of terminology that should be clarified here. The first concerns the terms "critical region" and "rejection region." Readers of other books might encounter either of the terms "critical region" or "rejection region" referring to either the set S_1 in Definition 9.1.4 or the set R in Definition 9.1.5. Those books generally define only one of the two terms. We choose to give the two sets S_1 and R different names because they are mathematically different objects. One, S_1, is a subset of the set of possible data vectors, while the other, R, is a subset of the set of possible values of a test statistic. Each has its use in different parts of the development of hypothesis testing. In most practical problems, tests are more easily expressed in terms of test statistics and rejection regions. For proving some theorems in Sec. 9.2, it is more convenient to define tests in terms of critical regions.

The final point of terminology concerns the terms "level of significance" and "size," as well as the term "level α_0 test." Some authors define level of significance (or significance level) for a test using a phrase such as "the probability of type I error" or "the probability that the data lie in the critical region when the null hypothesis is true." If the null hypothesis is simple, these phrases are easily understood, and they match what we defined as the size of the test in such cases. On the other hand, if the null hypothesis is composite, such phrases are ill-defined. For each $\theta \in \Omega_0$, there will usually be a different probability that the test rejects H_0. Which, if any, is the level of significance? We have defined the size of a test to be the supremum of all of these probabilities. We have said that the test "has level of significance α_0" if the size is less than or equal to α_0. This means that a test has one size but many levels of significance. Every number from the size up to 1 is a level of significance. There is a sound reason for distinguishing the concepts of size and level of significance. In Example 9.1.9, the investigator wants to constrain the probability of type I error to be less than 0.1. The test statistic Y has a discrete distribution, and we saw that no test with size 0.1 is available. In that example, the investigator needed to choose a test whose size was 0.0473. This test still has level of significance 0.1 and is a level 0.1 test, despite having a different size. There are other more complicated situations in which one can construct a test δ that satisfies Eq. (9.1.6), that is, it has level of significance α_0, but for which it is not possible (without sophisticated numerical methods) to compute the actual size. An investigator who insists on using a particular level of significance α_0 can use such a test, and call it a level α_0 test, without being able to compute its size exactly. The most common example of this latter situation is one in which we wish to test hypotheses concerning two parameters simultaneously. For example, let $\theta = (\theta_1, \theta_2)$, and suppose that we wish to test the hypotheses

$$H_0 : \theta_1 = 0 \text{ and } \theta_2 = 1 \quad \text{versus} \quad H_1 : \theta_1 \neq 0 \text{ or } \theta_2 \neq 1 \text{ or both.} \tag{9.1.26}$$

The following result gives a way to contruct a level α_0 test of H_0.

Theorem 9.1.5 For $i = 1, \ldots, n$, let $H_{0,i}$ be a null hypothesis, and let δ_i be a level $\alpha_{0,i}$ test of $H_{0,i}$. Define the combined null hypothesis H_0 that all of $H_{0,1}, \ldots, H_{0,n}$ are simultaneously true. Let δ be the test that rejects H_0 if at least one of $\delta_1, \ldots, \delta_n$ rejects its corresponding null hypothesis. Then δ is a level $\sum_{i=1}^{n} \alpha_{0,i}$ test of H_0.

Proof For $i = 1, \ldots, n$, let A_i be the event that δ_i rejects $H_{0,i}$. Apply Theorem 1.5.8. ■

To test H_0 in (9.1.26), find two tests δ_1 and δ_2 such that δ_1 is a test with size $\alpha_0/2$ for testing $\theta_1 = 0$ versus $\theta_1 \neq 0$ and δ_2 is a test with size $\alpha_0/2$ for testing $\theta_2 = 1$ versus $\theta_2 \neq 1$. Let δ be the test that rejects H_0 if either δ_1 rejects $\theta_1 = 0$ or δ_2 rejects $\theta_2 = 1$ or both. Theorem 9.1.5 says that δ is a level α_0 test of H_0 versus H_1, but its exact size requires us to be able to calculate the probability that both δ_1 and δ_2 simultaneously reject their corresponding null hypotheses. Such a calculation is often intractable.

Finally, our definition of level of significance matches nicely with the use of p-values, as pointed out immediately after Definition 9.1.9.

Summary

Hypothesis testing is the problem of deciding whether θ lies in a particular subset Ω_0 of the parameter space or in its complement Ω_1. The statement that $\theta \in \Omega_0$ is called the null hypothesis and is denoted by H_0. The alternative hypothesis is the statement $H_1 : \theta \in \Omega_1$. If S is the set of all possible data values (vectors) that we might observe, a subset $S_1 \subset S$ is called the critical region of a test of H_0 versus H_1 if we choose to reject H_0 whenever the observed data $\boldsymbol{X}$ are in S_1 and not reject H_0 whenever $\boldsymbol{X} \notin S_1$. The power function of this test δ is $\pi(\theta|\delta) = \Pr(\boldsymbol{X} \in S_1|\theta)$. The size of the test δ is $\sup_{\theta \in \Omega_0} \pi(\theta|\delta)$. A test is said to be a level α_0 test if its size is at most α_0. The null hypothesis H_0 is simple if Ω_0 is a set with only one point; otherwise, H_0 is composite. Similarly, H_1 is simple if Ω_1 has a single point, and H_1 is composite otherwise. A type I error is rejecting H_0 when it is true. A type II error is not rejecting H_0 when it is false.

Hypothesis tests are typically constructed by using a test statistic T. The null hypothesis is rejected if T lies in some interval or if T lies outside of some interval. The interval is chosen to make the test have a desired significance level. The p-value is a more informative way to report the results of a test. The p-value can be computed easily whenever our test has the form "reject H_0 if $T \geq c$" for some statistic T. The p-value when $T = t$ is observed equals $\sup_{\theta \in \Omega_0} \Pr(T \geq t|\theta)$. We also showed how a confidence set can be considered as a way of reporting the results of a test of hypotheses. A coefficient $1 - \alpha_0$ confidence set for θ is the set of all $\theta_0 \in \Omega$, such that we would not reject $H_0 : \theta = \theta_0$ using a level α_0 test. These confidence sets are intervals when we test hypotheses about a one-dimensional parameter or a one-dimensional function of the parameter.

Exercises

1. Let X have the exponential distribution with parameter β. Suppose that we wish to test the hypotheses $H_0: \beta \ge 1$ versus $H_1: \beta < 1$. Consider the test procedure δ that rejects H_0 if $X \ge 1$.

a. Determine the power function of the test.

b. Compute the size of the test.

2. Suppose that $X_1, \dots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, and that the following hypotheses are to be tested:

$$H_0: \quad \theta \ge 2,$$
$$H_1: \quad \theta < 2.$$

Let $Y_n = \max\{X_1, \dots, X_n\}$, and consider a test procedure such that the critical region contains all the outcomes for which $Y_n \le 1.5$.

a. Determine the power function of the test.

b. Determine the size of the test.

3. Suppose that the proportion p of defective items in a large population of items is unknown, and that it is desired to test the following hypotheses:

$$H_0: \quad p = 0.2,$$
$$H_1: \quad p \ne 0.2.$$

Suppose also that a random sample of 20 items is drawn from the population. Let Y denote the number of defective items in the sample, and consider a test procedure δ such that the critical region contains all the outcomes for which either $Y \ge 7$ or $Y \le 1$.

a. Determine the value of the power function $\pi(p|\delta)$ at the points $p = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9,$ and 1; sketch the power function.

b. Determine the size of the test.

4. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance 1. Suppose also that μ_0 is a certain specified number, and that the following hypotheses are to be tested:

$$H_0: \quad \mu = \mu_0,$$
$$H_1: \quad \mu \ne \mu_0.$$

Finally, suppose that the sample size n is 25, and consider a test procedure such that H_0 is to be rejected if $|\overline{X}_n - \mu_0| \ge c$. Determine the value of c such that the size of the test will be 0.05.

5. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Classify each of the following hypotheses as either simple or composite:

a. H_0: $\mu = 0$ and $\sigma = 1$

b. H_0: $\mu > 3$ and $\sigma < 1$

c. H_0: $\mu = -2$ and $\sigma^2 < 5$

d. H_0: $\mu = 0$

6. Suppose that a single observation X is to be taken from the uniform distribution on the interval $\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$, and suppose that the following hypotheses are to be tested:

$$H_0: \quad \theta \le 3,$$
$$H_1: \quad \theta \ge 4.$$

Construct a test procedure δ for which the power function has the following values: $\pi(\theta|\delta) = 0$ for $\theta \le 3$ and $\pi(\theta|\delta) = 1$ for $\theta \ge 4$.

7. Return to the situation described in Example 9.1.7. Consider a different test δ^* that rejects H_0 if $Y_n \le 2.9$ or $Y_n \ge 4.5$. Let δ be the test described in Example 9.1.7.

a. Prove that $\pi(\theta|\delta^*) = \pi(\theta|\delta)$ for all $\theta \le 4$.

b. Prove that $\pi(\theta|\delta^*) < \pi(\theta|\delta)$ for all $\theta > 4$.

c. Which of the two tests seems better for testing the hypotheses (9.1.8)?

8. Assume that $X_1, \dots, X_n$ are i.i.d. with the normal distribution that has mean μ and variance 1. Suppose that we wish to test the hypotheses

$$H_0: \quad \mu \le \mu_0,$$
$$H_1: \quad \mu > \mu_0.$$

Consider the test that rejects H_0 if $Z \ge c$, where Z is defined in Eq. (9.1.10).

a. Show that $\Pr(Z \ge c|\mu)$ is an increasing function of μ.

b. Find c to make the test have size α_0.

9. Assume that $X_1, \dots, X_n$ are i.i.d. with the normal distribution that has mean μ and variance 1. Suppose that we wish to test the hypotheses

$$H_0: \quad \mu \ge \mu_0,$$
$$H_1: \quad \mu < \mu_0.$$

Find a test statistic T such that, for every c, the test δ_c that rejects H_0 when $T \ge c$ has power function $\pi(\mu|\delta_c)$ that is decreasing in μ.

10. In Exercise 8, assume that $Z = z$ is observed. Find a formula for the p-value.

11. Assume that $X_1, \dots, X_9$ are i.i.d. having the Bernoulli distribution with parameter p. Suppose that we wish to test the hypotheses

$$H_0: \quad p = 0.4,$$
$$H_1: \quad p \ne 0.4.$$

Let $Y = \sum_{i=1}^{9} X_i$.

a. Find c_1 and c_2 such that

$$\Pr(Y \le c_1 | p = 0.4) + \Pr(Y \ge c_2 | p = 0.4)$$

is as close as possible to 0.1 without being larger than 0.1.

b. Let δ be the test that rejects H_0 if either $Y \le c_1$ or $Y \ge c_2$. What is the size of the test δ_c?

c. Draw a graph of the power function of δ_c.

12. Consider a single observation X from a Cauchy distribution centered at θ. That is, the p.d.f. of X is

$$f(x|\theta) = \frac{1}{\pi[1 + (x - \theta)^2]}, \quad \text{for } -\infty < x < \infty.$$

Suppose that we wish to test the hypotheses

$$H_0: \quad \theta \le \theta_0,$$
$$H_1: \quad \theta > \theta_0.$$

Let δ_c be the test that rejects H_0 if $X \ge c$.

a. Show that $\pi(\theta|\delta_c)$ is an increasing function of θ.

b. Find c to make δ_c have size 0.05.

c. If $X = x$ is observed, find a formula for the p-value.

13. Let X have the Poisson distribution with mean θ. Suppose that we wish to test the hypotheses

$$H_0: \quad \theta \le 1.0,$$
$$H_1: \quad \theta > 1.0.$$

Let δ_c be the test that rejects H_0 if $X \ge c$. Find c to make the size of δ_c as close as possible to 0.1 without being larger than 0.1.

14. Let $X_1, \dots, X_n$ be i.i.d. with the exponential distribution with parameter θ. Suppose that we wish to test the hypotheses

$$H_0: \quad \theta \ge \theta_0,$$
$$H_1: \quad \theta < \theta_0.$$

Let $X = \sum_{i=1}^{n} X_i$. Let δ_c be the test that rejects H_0 if $X \ge c$.

a. Show that $\pi(\theta|\delta_c)$ is a decreasing function of θ.

b. Find c in order to make δ_c have size α_0.

c. Let $\theta_0 = 2$, $n = 1$, and $\alpha_0 = 0.1$. Find the precise form of the test δ_c and sketch its power function.

15. Let X have the uniform distribution on the interval $[0, \theta]$, and suppose that we wish to test the hypotheses

$$H_0: \quad \theta \le 1,$$
$$H_1: \quad \theta > 1.$$

We shall consider test procedures of the form "reject H_0 if $X \ge c$." For each possible value x of X, find the p-value if $X = x$ is observed.

16. Consider the confidence interval found in Exercise 5 in Sec. 8.5. Find the collection of hypothesis tests that are equivalent to this interval. That is, for each $c > 0$, find a test δ_c of the null hypothesis $H_{0,c} : \sigma^2 = c$ versus some alternative such that δ_c rejects $H_{0,c}$ if and only if c is not in the interval. Write the test in terms of a test statistic $T = r(\boldsymbol{X})$ being in or out of some nonrandom interval that depends on c.

17. Let $X_1, \dots, X_n$ be i.i.d. with a Bernoulli distribution that has parameter p. Let $Y = \sum_{i=1}^{n} X_i$. We wish to find a coefficient γ confidence interval for p of the form $(q(y), 1)$. Prove that, if $Y = y$ is observed, then $q(y)$ should be chosen to be the smallest value p_0 such that $\Pr(Y \ge y | p = p_0) \ge 1 - \gamma$.

18. Consider the situation described immediately before Eq. (9.1.12). Prove that the expression (9.1.12) equals the smallest α_0 such that we would reject H_0 at level of significance α_0.

19. Return to the situation described in Example 9.1.17. Suppose that we wish to test the hypotheses

$$\begin{aligned} H_0: &\quad \mu \ge \mu_0, \\ H_1: &\quad \mu < \mu_0 \end{aligned} \tag{9.1.27}$$

at level α_0. It makes sense to reject H_0 if $\overline{X}_n$ is small. Construct a one-sided coefficient $1 - \alpha_0$ confidence interval for μ such that we can reject H_0 if μ_0 is not in the interval. Make sure that the test formed in this way rejects H_0 if $\overline{X}_n$ is small.

20. Prove Theorem 9.1.3.

21. Return to the situations described in Example 9.1.17 and Exercise 19. We wish to compare what might happen if we switch the null and alternative hypotheses. That is, we want to compare the results of testing the hypotheses in (9.1.22) at level α_0 to the results of testing the hypotheses in (9.1.27) at level α_0.

a. Let $\alpha_0 < 0.5$. Prove that there are no possible data sets such that we would reject both of the null hypotheses simultaneously. That is, for every possible $\overline{X}_n$ and σ', we must fail to reject at least one of the two null hypotheses.

b. Let $\alpha_0 < 0.5$. Prove that there are data sets that would lead to failing to reject both null hypotheses. Also prove that there are data sets that would lead to rejecting each of the null hypotheses while failing to reject the other.

c. Let $\alpha_0 > 0.5$. Prove that there are data sets that would lead to rejecting both null hypotheses.

★ 9.2 Testing Simple Hypotheses

The simplest hypothesis-testing situation is that in which there are only two possible values of the parameter. In such cases, it is possible to identify a collection of test procedures that have certain optimal properties.

Introduction

Example 9.2.1 Service Times in a Queue. In Example 3.7.5, we modeled the service times $\boldsymbol{X} = (X_1, \dots, X_n)$ of n customers in a queue as having the joint distribution with joint p.d.f.

$$f_1(\boldsymbol{x}) = \begin{cases} \dfrac{2(n!)}{\left(2 + \sum_{i=1}^{n} x_i\right)^{n+1}} & \text{for all } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{9.2.1}$$

Suppose that a service manager is not sure how well this joint distribution describes the service times. As an alternative, she proposes to model the service times as a random sample of exponential random variables with parameter 1/2. This model says that the joint p.d.f. is

$$f_0(\boldsymbol{x}) = \begin{cases} \dfrac{1}{2^n} \exp\left(-\dfrac{1}{2}\sum_{i=1}^{n} x_i\right) & \text{for all } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{9.2.2}$$

For illustration, Fig. 9.5 shows both of these p.d.f.'s for the case of $n = 1$. If the manager observes several service times, how can she test which of the two distributions appears to describe the data? ◀

In this section, we shall consider problems of testing hypotheses in which a vector of observations comes from one of two possible joint distributions, and the statistician must decide from which distribution the vector actually came. In many problems, each of the two joint distributions is actually the distribution of a random sample from a univariate distribution. However, nothing that we present in this section will depend on whether or not the observations form a random sample. In Example 9.2.1, one of the joint distributions is that of a random sample, but the other is not. In problems of this type, the parameter space Ω contains exactly two points, and both the null hypothesis and the alternative hypothesis are simple.

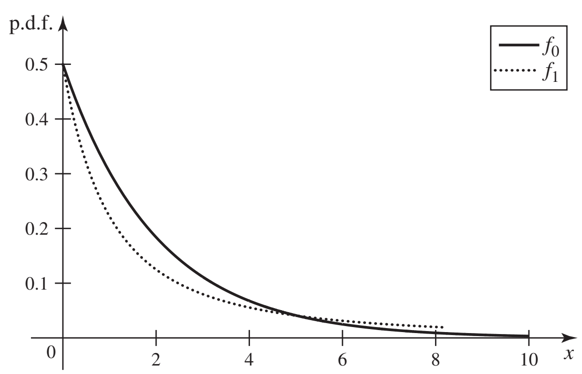


Figure 9.5 Graphs of the two competing p.d.f.'s in Example 9.2.1 with $n = 1$.

Specifically, we shall assume that the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ comes from a distribution for which the joint p.d.f., p.f., or p.f./p.d.f. is either $f_0(\boldsymbol{x})$ or $f_1(\boldsymbol{x})$. To correspond with notation earlier and later in the book, we can introduce a parameter space $\Omega = \{\theta_0, \theta_1\}$ and let $\theta = \theta_i$ stand for the case in which the data have p.d.f., p.f., or p.f./p.d.f. $f_i(\boldsymbol{x})$ for $i = 0, 1$. We are then interested in testing the following simple hypotheses:

$$\begin{aligned} H_0:&\quad \theta = \theta_0, \\ H_1:&\quad \theta = \theta_1. \end{aligned} \tag{9.2.3}$$

In this case, $\Omega_0 = \{\theta_0\}$ and $\Omega_1 = \{\theta_1\}$ are both singleton sets.

For the special case in which $\boldsymbol{X}$ is a random sample from a distribution with univariate p.d.f. or p.f. $f(x|\theta)$, we then have, for $i = 0$ or $i = 1$,

$$f_i(\boldsymbol{x}) = f(x_1|\theta_i) f(x_2|\theta_i) \cdots f(x_n|\theta_i).$$

The Two Types of Errors

When a test of the hypotheses (9.2.3) is being carried out, we have special notation for the probabilities of type I and type II errors. For each test procedure δ, we shall let $\alpha(\delta)$ denote the probability of an error of type I and shall let $\beta(\delta)$ denote the probability of an error of type II. Thus,

$$\begin{aligned} \alpha(\delta) &= \Pr(\text{Rejecting } H_0|\theta = \theta_0), \\ \beta(\delta) &= \Pr(\text{Not Rejecting } H_0|\theta = \theta_1). \end{aligned}$$

Example 9.2.2

Service Times in a Queue. The manager in Example 9.2.1 looks at the two p.d.f.'s in Fig. 9.5 and decides that f_1 gives higher probability to large service times than does f_0. So she decides to reject $H_0 : \theta = \theta_0$ if the service times are large. Specifically, suppose that she observes $n = 1$ service time, X_1. The test δ that she chooses rejects H_0 if $X_1 \geq 4$. The two error probabilities can be calculated from the two different possible distributions of X_1. Given $\theta = \theta_0$, X_1 has the exponential distribution with parameter 0.5. The c.d.f. of this distribution is $F_0(x) = 1 - \exp(-0.5x)$ for $x \geq 0$. The type I error probability is the probability that $X_1 \geq 4$, which equals $\alpha(\delta) = 0.135$. Given $\theta = \theta_1$, the distribution of X_1 has the p.d.f. $2/(2 + x_1)^2$ for $x_1 \geq 0$. The c.d.f. is then $F_1(x) = 1 - 2/(2 + x)$, for $x \geq 0$. The type II error probability is $\beta(\delta) = \Pr(X_1 < 4) = F_1(4) = 0.667$. ◀

It is desirable to find a test procedure for which the probabilities $\alpha(\delta)$ and $\beta(\delta)$ of the two types of error will be small. For a given sample size, it is typically not possible to find a test procedure for which both $\alpha(\delta)$ and $\beta(\delta)$ will be arbitrarily small. Therefore, we shall now show how to construct a procedure for which the value of a specific linear combination of α and β will be minimized.

Optimal Tests

Minimizing a Linear Combination Suppose that a and b are specified positive constants, and it is desired to find a procedure δ for which $a\alpha(\delta) + b\beta(\delta)$ will be a minimum. Theorem 9.2.1 shows that a procedure that is optimal in this sense has a very simple form. In Sec. 9.8, we shall give a rationale for choosing a test to minimize a linear combination of the error probabilities.

Theorem 9.2.1 Let δ^* denote a test procedure such that the hypothesis H_0 is not rejected if $af_0(\boldsymbol{x}) > bf_1(\boldsymbol{x})$ and the hypothesis H_0 is rejected if $af_0(\boldsymbol{x}) < bf_1(\boldsymbol{x})$. The null hypothesis H_0 can be either rejected or not if $af_0(\boldsymbol{x}) = bf_1(\boldsymbol{x})$. Then for every other test procedure δ,

$$a\alpha(\delta^*) + b\beta(\delta^*) \le a\alpha(\delta) + b\beta(\delta). \tag{9.2.4}$$

Proof For convenience, we shall present the proof for a problem in which the random sample $X_1, \dots, X_n$ is drawn from a discrete distribution. In this case, $f_i(\boldsymbol{x})$ represents the joint p.f. of the observations in the sample when H_i is true ($i = 0, 1$). If the sample comes from a continuous distribution, in which case $f_i(\boldsymbol{x})$ is a joint p.d.f., then each of the sums that will appear in this proof should be replaced by an n-dimensional integral.

If we let S_1 denote the critical region of an arbitrary test procedure δ, then S_1 contains every sample outcome $\boldsymbol{x}$ for which δ specifies that H_0 should be rejected, and $S_0 = S_1^c$ contains every outcome $\boldsymbol{x}$ for which H_0 should not be rejected. Therefore,

$$\begin{aligned} a\alpha(\delta) + b\beta(\delta) &= a \sum_{\boldsymbol{x}\in S_1} f_0(\boldsymbol{x}) + b \sum_{\boldsymbol{x}\in S_0} f_1(\boldsymbol{x}) \\ &= a \sum_{\boldsymbol{x}\in S_1} f_0(\boldsymbol{x}) + b\left[1 - \sum_{\boldsymbol{x}\in S_1} f_1(\boldsymbol{x})\right] \\ &= b + \sum_{\boldsymbol{x}\in S_1} [af_0(\boldsymbol{x}) - bf_1(\boldsymbol{x})]. \end{aligned} \tag{9.2.5}$$

It follows from Eq. (9.2.5) that the value of the linear combination $a\alpha(\delta) + b\beta(\delta)$ will be a minimum if the critical region S_1 is chosen so that the value of the final summation in Eq. (9.2.5) is a minimum. Furthermore, the value of this summation will be a minimum if the summation includes every point $\boldsymbol{x}$ for which $af_0(\boldsymbol{x}) - bf_1(\boldsymbol{x}) < 0$ and includes no point $\boldsymbol{x}$ for which $af_0(\boldsymbol{x}) - bf_1(\boldsymbol{x}) > 0$. In other words, $a\alpha(\delta) + b\beta(\delta)$ will be a minimum if the critical region S_1 is chosen to include every point $\boldsymbol{x}$ such that $af_0(\boldsymbol{x}) < bf_1(\boldsymbol{x})$ and exclude every point $\boldsymbol{x}$ such that this inequality is reversed. If $af_0(\boldsymbol{x}) = bf_1(\boldsymbol{x})$ for some point $\boldsymbol{x}$, then it is irrelevant whether or not $\boldsymbol{x}$ is included in S_1, because the corresponding term would contribute zero to the final summation in Eq. (9.2.5). The critical region described above corresponds to the test procedure δ^* defined in the statement of the theorem. ■

The ratio $f_1(\boldsymbol{x})/f_0(\boldsymbol{x})$ is sometimes called the *likelihood ratio* of the sample. It is related to, but not the same as, the likelihood ratio statistic from Definition 9.1.11. In the present context, the likelihood ratio statistic $\Lambda(\boldsymbol{x})$ would equal $f_0(\boldsymbol{x})/\max\{f_0(\boldsymbol{x}), f_1(\boldsymbol{x})\}$. In particular, the likelihood ratio $f_1(\boldsymbol{x})/f_0(\boldsymbol{x})$ is large when $\Lambda(\boldsymbol{x})$ is small, and vice versa. In fact,

$$\Lambda(\boldsymbol{x}) = \begin{cases} \left(\dfrac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})}\right)^{-1} & \text{if } f_0(\boldsymbol{x}) \le f_1(\boldsymbol{x}) \\ 1 & \text{otherwise.} \end{cases}$$

The important point to remember about this confusing choice of names is the following: The theoretical justification for tests based on the likelihood ratio defined here (provided in Theorems 9.2.1 and 9.2.2) is the rationale for expecting the likelihood ratio tests of Defintion 9.1.11 to be sensible.

When $a, b > 0$, Theorem 9.2.1 can be reworded as follows.

Corollary 9.2.1 Assume the conditions of Theorem 9.2.1, and assume that $a > 0$ and $b > 0$. Then the test δ for which the value of $a\alpha(\delta) + b\beta(\delta)$ is a minimum rejects H_0 when the

likelihood ratio exceeds a/b and does not reject H_0 when the likelihood ratio is less than a/b. ■

Example 9.2.3 Service Times in a Queue. Instead of rejecting H_0 if $X_1 \geq 4$ in Example 9.2.2, the manager could apply Theorem 9.2.1. She must choose two numbers a and b to balance the two types of error. Suppose that she chooses them to be equal to each other. Then the test will be to reject H_0 if $f_1(x_1)/f_0(x_1) > 1$. That is, if

$$\frac{4}{(2+x_1)^2} \exp\left(\frac{x_1}{2}\right) > 1. \tag{9.2.6}$$

At $x_1 = 0$ the left side of Eq. (9.2.6) equals 1, and it decreases until $x_1 = 2$ and then increases ever after. Hence, Eq. (9.2.6) holds for all values of $x_1 > c$ where c is the unique strictly positive value where the left side of Eq. (9.2.6) equals 1. By numerical approximation, we find that this value is $x_1 = 5.025725$. The type I and type II error probabilities for the test δ^* that rejects H_0 if $X_1 > 5.025725$ are

$$\alpha(\delta^*) = 1 - F_0(5.025725) = \exp(-2.513) = 0.081,$$

$$\beta(\delta^*) = F_1(5.025725) = 1 - \frac{2}{7.026} = 0.715.$$

The sum of these error probabilities is 0.796. By comparison, the sum of the two error probabilities in Example 9.2.2 is 0.802, slightly higher. ◄

Minimizing the Probability of an Error of Type II Next, suppose that the probability $\alpha(\delta)$ of an error of type I is not permitted to be greater than a specified level of significance, and it is desired to find a procedure δ for which $\beta(\delta)$ will be a minimum. In this problem, we can apply the following result, which is closely related to Theorem 9.2.1 and is known as the *Nayman-Pearson lemma* in honor of the statisticians J. Neyman and E. S. Pearson, who developed these ideas in 1933.

Theorem 9.2.2 Nayman-Pearson lemma. Suppose that δ' is a test procedure that has the following form for some constant $k > 0$: The hypothesis H_0 is not rejected if $f_1(\boldsymbol{x}) < kf_0(\boldsymbol{x})$ and the hypothesis H_0 is rejected if $f_1(\boldsymbol{x}) > kf_0(\boldsymbol{x})$. The null hypothesis H_0 can be either rejected or not if $f_1(\boldsymbol{x}) = kf_0(\boldsymbol{x})$. If δ is another test procedure such that $\alpha(\delta) \leq \alpha(\delta')$, then it follows that $\beta(\delta) \geq \beta(\delta')$. Furthermore, if $\alpha(\delta) < \alpha(\delta')$, then $\beta(\delta) > \beta(\delta')$.

Proof From the description of the procedure δ' and from Theorem 9.2.1, it follows that for every test procedure δ,

$$k\alpha(\delta') + \beta(\delta') \leq k\alpha(\delta) + \beta(\delta). \tag{9.2.7}$$

If $\alpha(\delta) \leq \alpha(\delta')$, then it follows from the relation (9.2.7) that $\beta(\delta) \geq \beta(\delta')$. Also, if $\alpha(\delta) < \alpha(\delta')$, then it follows that $\beta(\delta) > \beta(\delta')$. ■

To illustrate the use of the Nayman-Pearson lemma, we shall suppose that a statistician wishes to use a test procedure for which $\alpha(\delta) = \alpha_0$ and $\beta(\delta)$ is a minimum. According to the lemma, she should try to find a value of k for which $\alpha(\delta') = \alpha_0$. The procedure δ' will then have the minimum possible value of $\beta(\delta)$. If the distribution from which the sample is taken is continuous, then it is usually (but not always) possible to find a value of k such that $\alpha(\delta')$ is equal to a specified value such as α_0. However, if the distribution from which the sample is taken is discrete, then it is typically not possible to choose k so that $\alpha(\delta')$ is equal to a specified value. These remarks are considered further in the following examples and in the exercises at the end of this section.

Example 9.2.4

Service Times in a Queue. In Example 9.2.3, the distribution of X_1 is continuous, and we can find a value k such that the test δ' that results from Theorem 9.2.2 has $\alpha(\delta') = 0.07$, say. The test δ^* in Example 9.2.3 has $\alpha(\delta^*) > 0.07$ and $k = 1$. We will need a larger value of k in order to get the type I error probability down to 0.07. As we noted in Example 9.2.3, the left side of Eq. (9.2.6) is increasing for $x_1 > 2$, and hence the set of x_1 values such that

$$\frac{4}{(2+x_1)^2}\exp\left(\frac{x_1}{2}\right) > k \tag{9.2.8}$$

will be an interval of the form (c, ∞) where c is the unique value that makes the left side of Eq. (9.2.8) equal to k. The resulting test will then have the form "reject H_0 if $X_1 \ge c$." At this point, we don't care any more about k because we just need to choose c to make sure that $\Pr(X_1 \ge c|\theta = \theta_0) = 0.07$. That is, we need $1 - F_0(c) = 0.07$. Recall that $F_0(c) = 1 - \exp(-0.5c)$, so $c = -2\log(0.07) = 5.318$. We can then compute $\beta(\delta') = F_1(5.318) = 0.727$. This test is very close to δ^* from Example 9.2.3. ◀

Example 9.2.5

Random Sample from a Normal Distribution. Suppose that $\boldsymbol{X} = (X_1, \dots, X_n)$ is a random sample from the normal distribution with unknown mean θ and known variance 1, and the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad \theta = 0,\\ H_1&: \quad \theta = 1. \end{aligned} \tag{9.2.9}$$

We shall begin by determining a test procedure for which $\beta(\delta)$ will be a minimum among all test procedures for which $\alpha(\delta) \le 0.05$.

When H_0 is true, the variables $X_1, \dots, X_n$ form a random sample from the standard normal distribution. When H_1 is true, these variables form a random sample from the normal distribution for which both the mean and the variance are 1. Therefore,

$$f_0(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \tag{9.2.10}$$

and

$$f_1(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - 1)^2\right]. \tag{9.2.11}$$

After some algebraic simplification, the likelihood ratio $f_1(\boldsymbol{x})/f_0(\boldsymbol{x})$ can be written in the form

$$\frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} = \exp\left[n\left(\overline{x}_n - \frac{1}{2}\right)\right]. \tag{9.2.12}$$

It now follows from Eq. (9.2.12) that rejecting the hypothesis H_0 when the likelihood ratio is greater than a specified positive constant k is equivalent to rejecting H_0 when the sample mean $\overline{x}_n$ is greater than $(1/2) + (1/n)\log k$.

Let $k' = (1/2) + (1/n)\log k$, and suppose that we can find a value of k' such that

$$\Pr\left(\overline{X}_n > k'|\theta = 0\right) = 0.05. \tag{9.2.13}$$

Then the procedure δ', which rejects H_0 when $\overline{X}_n > k'$, will satisfy $\alpha(\delta') = 0.05$. Furthermore, by the Nayman-Pearson lemma, δ' will be an optimal procedure in the sense of minimizing the value of $\beta(\delta)$ among all procedures for which $\alpha(\delta) \le 0.05$.

It is easy to see that the value of k' that satisfies Eq. (9.2.13) must be the 0.95 quantile of the distribution of $\overline{X}_n$ given $\theta = 0$. When $\theta = 0$, the distribution of $\overline{X}_n$ is the normal distribution with mean 0 and variance $1/n$. Therefore, its 0.95 quantile is $0 + \Phi^{-1}(0.95)n^{-1/2}$, where Φ^{-1} is the standard normal quantile function. From a table of the standard normal distribution, it is found that the 0.95 quantile of the standard normal distribution is 1.645, so $k' = 1.645n^{-1/2}$.

In summary, among all test procedures for which $\alpha(\delta) \le 0.05$, the procedure that rejects H_0 when $\overline{X}_n > 1.645n^{-1/2}$ has the smallest probability of type II error.

Next, we shall determine the probability $\beta(\delta')$ of an error of type II for this procedure δ'. Since $\beta(\delta')$ is the probability of not rejecting H_0 when H_1 is true,

$$\beta(\delta') = \Pr(\overline{X}_n < 1.645n^{-1/2} | \theta = 1). \tag{9.2.14}$$

When $\theta = 1$, the distribution of $\overline{X}_n$ is the normal distribution with mean 1 and variance $1/n$. The probability in Eq. (9.2.14) can then be written as

$$\beta(\delta') = \Phi\left(\frac{1.645n^{-1/2} - 1}{n^{-1/2}}\right) = \Phi(1.645 - n^{1/2}). \tag{9.2.15}$$

For instance, when $n = 9$, it is found from a table of the standard normal distribution that

$$\beta(\delta') = \Phi(-1.355) = 1 - \Phi(1.355) = 0.0877.$$

Finally, for this same random sample and the same hypotheses (9.2.9), we shall determine the test procedure δ_0 for which the value of $2\alpha(\delta) + \beta(\delta)$ is a minimum, and we shall calculate the value of $2\alpha(\delta_0) + \beta(\delta_0)$ when $n = 9$.

It follows from Theorem 9.2.1 that the procedure δ_0 for which $2\alpha(\delta) + \beta(\delta)$ is a minimum rejects H_0 when the likelihood ratio is greater than 2. By Eq. (9.2.12), this procedure is equivalent to rejecting H_0 when $\overline{X}_n > (1/2) + (1/n) \log 2$. Thus, when $n = 9$, the optimal procedure δ_0 rejects H_0 when $\overline{X}_n > 0.577$. For this procedure we then have

$$\alpha(\delta_0) = \Pr(\overline{X}_n > 0.577 | \theta = 0) \tag{9.2.16}$$

and

$$\beta(\delta_0) = \Pr(\overline{X}_n < 0.577 | \theta = 1). \tag{9.2.17}$$

Since $\overline{X}_n$ has the normal distribution with mean θ and variance $1/n$, we have

$$\alpha(\delta_0) = 1 - \Phi\left(\frac{0.577 - 0}{1/3}\right) = 1 - \Phi(1.731) = 0.0417$$

and

$$\beta(\delta_0) = \Phi\left(\frac{0.577 - 1}{1/3}\right) = \Phi(-1.269) = 0.1022.$$

The minimum value of $2\alpha(\delta) + \beta(\delta)$ is therefore

$$2\alpha(\delta_0) + \beta(\delta_0) = 2(0.0417) + (0.1022) = 0.1856. \qquad \blacktriangleleft$$

Example 9.2.6

Sampling from a Bernoulli Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p, and the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad p = 0.2, \\ H_1&: \quad p = 0.4. \end{aligned} \tag{9.2.18}$$

It is desired to find a test procedure for which $\alpha(\delta) = 0.05$ and $\beta(\delta)$ is a minimum.

In this example, each observed value x_i must be either 0 or 1. If we let $y = \sum_{i=1}^{n} x_i$, then the joint p.f. of $X_1, \ldots, X_n$ when $p = 0.2$ is

$$f_0(\boldsymbol{x}) = (0.2)^y (0.8)^{n-y} \tag{9.2.19}$$

and the joint p.f. when $p = 0.4$ is

$$f_1(\boldsymbol{x}) = (0.4)^y (0.6)^{n-y}. \tag{9.2.20}$$

Hence, the likelihood ratio is

$$\frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} = \left(\frac{3}{4}\right)^n \left(\frac{8}{3}\right)^y. \tag{9.2.21}$$

It follows that rejecting H_0 when the likelihood ratio is greater than a specified positive constant k is equivalent to rejecting H_0 when y is greater than k', where

$$k' = \frac{\log k + n \log(4/3)}{\log(8/3)}. \tag{9.2.22}$$

To find a test procedure for which $\alpha(\delta) = 0.05$ and $\beta(\delta)$ is a minimum, we use the Nayman-Pearson lemma. If we let $Y = \sum_{i=1}^{n} X_i$, we should try to find a value of k' such that

$$\Pr(Y > k' | p = 0.2) = 0.05. \tag{9.2.23}$$

When the hypothesis H_0 is true, the random variable Y has the binomial distribution with parameters n and $p = 0.2$. However, because of the discreteness of this distribution, it generally will not be possible to find a value of k' for which Eq. (9.2.23) is satisfied. For example, suppose that $n = 10$. Then it is found from a table of the binomial distribution that $\Pr(Y > 4 | p = 0.2) = 0.0328$ and also $\Pr(Y > 3 | p = 0.2) = 0.1209$. Therefore, there is no critical region of the desired form for which $\alpha(\delta) = 0.05$. If it is desired to use a level 0.05 test δ based on the likelihood ratio as specified by the Nayman-Pearson lemma, then one must reject H_0 when $Y > 4$ and $\alpha(\delta) = 0.0328$. ◄

Randomized Tests

It has been emphasized by some statisticians that $\alpha(\delta)$ can be made exactly 0.05 in Example 9.2.6 if a *randomized* test procedure is used. Such a procedure is described as follows: When the rejection region of the test procedure contains all values of y greater than 4, we found in Example 9.2.6 that the size of the test is $\alpha(\delta) = 0.0328$. Also, when the point $y = 4$ is added to this rejection region, the value of $\alpha(\delta)$ jumps to 0.1209. Suppose, however, that instead of choosing between including the point $y = 4$ in the rejection region and excluding that point, we use an auxiliary randomization to decide whether or not to reject H_0 when $y = 4$. For example, we may toss a coin or spin a wheel to arrive at this decision. Then, by choosing appropriate probabilities to be used in this randomization, we can make $\alpha(\delta)$ exactly 0.05.

Specifically, consider the following test procedure: The hypothesis H_0 is rejected if $y > 4$, and H_0 is not rejected if $y < 4$. However, if $y = 4$, then an auxiliary randomization is carried out in which H_0 will be rejected with probability 0.195, and H_0 will not be rejected with probability 0.805. The size $\alpha(\delta)$ of this test will then be

$$\begin{aligned}\alpha(\delta) &= \Pr(Y > 4 | p = 0.2) + (0.195)\Pr(Y = 4 | p = 0.2)\\ &= 0.0328 + (0.195)(0.0881) = 0.05.\end{aligned} \tag{9.2.24}$$

Randomized tests do not seem to have any place in practical applications of statistics. It does not seem reasonable for a statistician to decide whether or not to reject a null hypothesis by tossing a coin or performing some other type of randomization for the sole purpose of obtaining a value of $\alpha(\delta)$ that is equal to some arbitrarily specified value such as 0.05. The main consideration for the statistician is to use a nonrandomized test procedure δ' having the form specified in the Nayman-Pearson lemma.

The proofs of Theorems 9.2.1 and 9.2.2 can be extended to find optimal tests among all tests regardless of whether they are randomized or nonrandomized. The optimal test in the extension of Theorem 9.2.2 has the same form as δ^* except that randomization is allowed whenever $f_1(\boldsymbol{x}) = kf_0(\boldsymbol{x})$. The only real need for randomized tests, in this book, will be the simplification that they provide for one step in the proof of Theorem 9.3.1 (page 562).

Furthermore, rather than fixing a specific size $\alpha(\delta)$ and trying to minimize $\beta(\delta)$, it might be more reasonable for the statistician to minimize a linear combination of the form $a\alpha(\delta) + b\beta(\delta)$. As we have seen in Theorem 9.2.1, such a minimization can always be achieved without recourse to an auxiliary randomization. In Sec. 9.9, we shall present another argument that indicates why it might be more reasonable to minimize a linear combination of the form $a\alpha(\delta) + b\beta(\delta)$ than to specify a value of $\alpha(\delta)$ and then minimize $\beta(\delta)$.

Summary

For the special case in which there are only two possible values, θ_0 and θ_1, for the parameter, we found a collection of procedures for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ that contains the optimal test procedure for each of the following criteria:

- Choose the test δ with the smallest value of $a\alpha(\delta) + b\beta(\delta)$.
- Among all tests δ with $\alpha(\delta) \le \alpha_0$, choose the test with the smallest value of $\beta(\delta)$.

Here, $\alpha(\delta) = \Pr(\text{Reject } H_0 | \theta = \theta_0)$ and $\beta(\delta) = \Pr(\text{Don't Reject } H_0 | \theta = \theta_1)$ are, respectively, the probabilities of type I and type II errors. The tests all have the following form for some positive constant k: reject H_0 if $f_0(\boldsymbol{x}) < kf_1(\boldsymbol{x})$, don't reject H_0 if $f_0(\boldsymbol{x}) > kf_1(\boldsymbol{x})$, and do either if $f_0(\boldsymbol{x}) = kf_1(\boldsymbol{x})$.

Exercises

1. Let $f_0(x)$ be the p.f. of the Bernoulli distribution with parameter 0.3, and let $f_1(x)$ be the p.f. of the Bernoulli distribution with parameter 0.6. Suppose that a single observation X is taken from a distribution for which the p.d.f. $f(x)$ is either $f_0(x)$ or $f_1(x)$, and the following simple hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad f(x) = f_0(x), \\ H_1&: \quad f(x) = f_1(x). \end{aligned}$$

Find the test procedure δ for which the value of $\alpha(\delta) + \beta(\delta)$ is a minimum.

2. Consider two p.d.f.'s $f_0(x)$ and $f_1(x)$ that are defined as follows:

$$f_0(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_1(x) = \begin{cases} 2x & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that a single observation X is taken from a distribution for which the p.d.f. $f(x)$ is either $f_0(x)$ or $f_1(x)$, and the following simple hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad f(x) = f_0(x), \\ H_1&: \quad f(x) = f_1(x). \end{aligned}$$

a. Describe a test procedure for which the value of $\alpha(\delta) + 2\beta(\delta)$ is a minimum.

b. Determine the minimum value of $\alpha(\delta) + 2\beta(\delta)$ attained by that procedure.

3. Consider again the conditions of Exercise 2, but suppose now that it is desired to find a test procedure for which the value of $3\alpha(\delta) + \beta(\delta)$ is a minimum.

a. Describe the procedure.

b. Determine the minimum value of $3\alpha(\delta) + \beta(\delta)$ attained by the procedure.

4. Consider again the conditions of Exercise 2, but suppose now that it is desired to find a test procedure for which $\alpha(\delta) \le 0.1$ and $\beta(\delta)$ is a minimum.

a. Describe the procedure.

b. Determine the minimum value of $\beta(\delta)$ attained by the procedure.

5. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean θ and known variance is 1, and the following hypotheses are to be tested:

$$H_0: \quad \theta = 3.5,$$
$$H_1: \quad \theta = 5.0.$$

a. Among all test procedures for which $\beta(\delta) \le 0.05$, describe a procedure for which $\alpha(\delta)$ is a minimum.

b. For $n = 4$, find the minimum value of $\alpha(\delta)$ attained by the procedure described in part (a).

6. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p. Let p_0 and p_1 be specified values such that $0 < p_1 < p_0 < 1$, and suppose that it is desired to test the following simple hypotheses:

$$H_0: \quad p = p_0,$$
$$H_1: \quad p = p_1.$$

a. Show that a test procedure for which $\alpha(\delta) + \beta(\delta)$ is a minimum rejects H_0 when $\overline{X}_n < c$.

b. Find the value of the constant c.

7. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with known mean μ and unknown variance σ^2, and the following simple hypotheses are to be tested:

$$H_0: \quad \sigma^2 = 2,$$
$$H_1: \quad \sigma^2 = 3.$$

a. Show that among all test procedures for which $\alpha(\delta) \le 0.05$, the value of $\beta(\delta)$ is minimized by a procedure that rejects H_0 when $\sum_{i=1}^{n}(X_i - \mu)^2 > c$.

b. For $n = 8$, find the value of the constant c that appears in part (a).

8. Suppose that a single observation X is taken from the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown, and the following simple hypotheses are to be tested:

$$H_0: \quad \theta = 1,$$
$$H_1: \quad \theta = 2.$$

a. Show that there exists a test procedure for which $\alpha(\delta) = 0$ and $\beta(\delta) < 1$.

b. Among all test procedures for which $\alpha(\delta) = 0$, find the one for which $\beta(\delta)$ is a minimum.

9. Suppose that a random sample $X_1, \ldots, X_n$ is drawn from the uniform distribution on the interval $[0, \theta]$, and consider again the problem of testing the simple hypotheses described in Exercise 8. Find the minimum value of $\beta(\delta)$ that can be attained among all test procedures for which $\alpha(\delta) = 0$.

10. Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with unknown mean λ. Let λ_0 and λ_1 be specified values such that $\lambda_1 > \lambda_0 > 0$, and suppose that it is desired to test the following simple hypotheses:

$$H_0: \quad \lambda = \lambda_0,$$
$$H_1: \quad \lambda = \lambda_1.$$

a. Show that the value of $\alpha(\delta) + \beta(\delta)$ is minimized by a test procedure which rejects H_0 when $\overline{X}_n > c$.

b. Find the value of c.

c. For $\lambda_0 = 1/4$, $\lambda_1 = 1/2$, and $n = 20$, determine the minimum value of $\alpha(\delta) + \beta(\delta)$ that can be attained.

11. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known standard deviation 2, and the following simple hypotheses are to be tested:

$$H_0: \quad \mu = -1,$$
$$H_1: \quad \mu = 1.$$

Determine the minimum value of $\alpha(\delta) + \beta(\delta)$ that can be attained for each of the following values of the sample size n:

a. $n = 1$ **b.** $n = 4$ **c.** $n = 16$ **d.** $n = 36$

12. Let $X_1, \ldots, X_n$ be a random sample from the exponential distribution with unknown parameter θ. Let $0 < \theta_0 < \theta_1$ be two possible values of the parameter. Suppose that we wish to test the following hypotheses:

$$H_0: \quad \theta = \theta_0,$$
$$H_1: \quad \theta = \theta_1.$$

For each $\alpha_0 \in (0, 1)$, show that among all tests δ satisfying $\alpha(\delta) \le \alpha_0$, the test with the smallest probability of type II error will reject H_0 if $\sum_{i=1}^{n} X_i < c$, where c is the α_0 quantile of the gamma distribution with parameters n and θ_0.

13. Consider the series of examples in this section concerning service times in a queue. Suppose that the manager observes two service times X_1 and X_2. It is easy to see that both $f_1(\boldsymbol{x})$ in (9.2.1) and $f_0(\boldsymbol{x})$ in (9.2.2) depend on the observed data only through the value $t = x_1 + x_2$ of the statistic $T = X_1 + X_2$. Hence, the tests from Theorems 9.2.1 and 9.2.2 both depend only on the value of T.

a. Using Theorem 9.2.1, determine the test procedure that minimizes the sum of the probabilities of type I and type II errors.

b. Suppose that $X_1 = 4$ and $X_2 = 3$ are observed. Perform the test in part (a) to see whether H_0 is rejected.

c. Prove that the distribution of T, given that H_0 is true, is the gamma distribution with parameters 2 and 1/2.

d. Using Theorem 9.2.2, determine the test procedure with level at most 0.01 that has minimum probability of type II error. *Hint:* It looks like you need to solve a system of nonlinear equations, but for a level 0.01 test, the equations collapse to a single simple equation.

e. Suppose that $X_1 = 4$ and $X_2 = 3$ are observed. Perform the test in part (d) to see whether H_0 is rejected.

★ 9.3 Uniformly Most Powerful Tests

When the null and/or alternative hypothesis is composite, we can still find a class of tests that has optimal properties in certain circumstances. In particular, the null and alternative hypotheses must be of the form $H_0 : \theta \le \theta_0$ and $H_1 : \theta > \theta_0$, or $H_0 : \theta \ge \theta_0$ and $H_1 : \theta < \theta_0$. In addition, the family of distributions of the data must have a property called "monotone likelihood ratio," which is defined in this section.

Definition of a Uniformly Most Powerful Test

Example 9.3.1

Service Times in a Queue. In Example 9.2.1, a manager was interested in testing which of two joint distributions described the service times in a queue that she was managing. Suppose, now, that instead of considering only two joint distributions, the manager wishes to consider all of the joint distributions that can be described by saying that the service times form a random sample from the exponential distribution with parameter θ conditional on θ. That is, for each possible rate $\theta > 0$, the manager is willing to consider the possibility that the service times are i.i.d. exponential random variables with parameter θ. In particular, the manager is interested in testing $H_0 : \theta \le 1/2$ versus $H_1 : \theta > 1/2$. For each $\theta' > 1/2$, the manager could use the methods of Sec. 9.2 to test the hypotheses $H_0' : \theta = 1/2$ versus $H_1' : \theta = \theta'$. She could obtain the level α_0 test with the smallest possible type II error probability when $\theta = \theta'$. But can she find a single level α_0 test that has the largest possible type II error probability simultaneously for all $\theta > 1/2$? And will that test have probability of type I error at most α_0 for all $\theta \le 1/2$? ◄

Consider a problem of testing hypotheses in which the random variables $\boldsymbol{X} = (X_1, \dots, X_n)$ form a random sample from a distribution for which either the p.d.f. or the p.f. is $f(x|\theta)$. We suppose that the value of the parameter θ is unknown but must lie in a specified parameter space Ω that is a subset of the real line. As usual, we shall suppose that Ω_0 and Ω_1 are disjoint subsets of Ω, and the hypotheses to be tested are

$$\begin{aligned} H_0&: \quad \theta \in \Omega_0, \\ H_1&: \quad \theta \in \Omega_1. \end{aligned} \tag{9.3.1}$$

We shall assume that the subset Ω_1 contains at least two distinct values of θ, in which case the alternative hypothesis H_1 is composite. The null hypothesis H_0 may be either simple or composite. Example 9.3.1 is of the type just described with $\Omega_0 = (0, 1/2]$ and $\Omega_1 = (1/2, \infty)$.

We shall also suppose that it is desired to test the hypotheses (9.3.1) at a specified level of significance α_0, where α_0 is a given number in the interval $0 < \alpha_0 < 1$. In other words, we shall consider only procedures in which $\Pr(\text{Rejecting } H_0|\theta) \le \alpha_0$ for every value of $\theta \in \Omega_0$. If $\pi(\theta|\delta)$ denotes the power function of a given test procedure δ, this requirement can be written simply as

$$\pi(\theta|\delta) \le \alpha_0 \quad \text{for } \theta \in \Omega_0. \tag{9.3.2}$$

Equivalently, if $\alpha(\delta)$ denotes the size of a test procedure δ, as defined by Eq. (9.1.7), then the requirement (9.3.2) can also be expressed by the relation

$$\alpha(\delta) \le \alpha_0. \tag{9.3.3}$$

Finally, among all test procedures that satisfy the requirement (9.3.3), we want to find one that has the smallest possible probability of type II error for every $\theta \in \Omega_1$. In terms of the power function, we want the value of $\pi(\theta|\delta)$ to be as large as possible for every value of $\theta \in \Omega_1$.

It may not be possible to satisfy this last criterion. If θ_1 and θ_2 are two different values of θ in Ω_1, then the test procedure for which the value of $\pi(\theta_1|\delta)$ is a maximum might be different from the test procedure for which the value of $\pi(\theta_2|\delta)$ is a maximum. In other words, there might be no single test procedure δ that maximizes the power function $\pi(\theta|\delta)$ simultaneously for every value of θ in Ω_1. In some problems, however, there will exist a test procedure that satisfies this criterion. Such a procedure, when it exists, is called a *uniformly most powerful* test, or, more briefly, a UMP test. The formal definition of a UMP test is as follows.

Definition 9.3.1 Uniformly Most Powerful (UMP) Test. A test procedure δ^* is a *uniformly most powerful (UMP) test* of the hypotheses (9.3.1) at the level of significance α_0 if $\alpha(\delta^*) \le \alpha_0$ and, for every other test procedure δ such that $\alpha(\delta) \le \alpha_0$, it is true that

$$\pi(\theta|\delta) \le \pi(\theta|\delta^*) \quad \text{for every value of } \theta \in \Omega_1. \tag{9.3.4}$$

In this section, we shall show that a UMP test exists in many problems in which the random sample comes from one of the standard families of distributions that we have been considering in this book.

Monotone Likelihood Ratio

Example 9.3.2 Service Times in a Queue. Suppose that the manager in Example 9.3.1 observes a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ of service times and tries to find the level α_0 test of $H_0' : \theta = 1/2$ versus $H_1' : \theta = \theta'$ that has the largest power at $\theta = \theta' > 1/2$. According to Exercise 12 in Sec. 9.2, the test will reject H_0' if $\sum_{i=1}^{n} X_i$ is less than the α_0 quantile of the gamma distribution with parameters n and $1/2$. This test is the same test regardless of which $\theta' > 1/2$ the manager considers. Hence, the test is UMP at the level of significance α_0 for testing $H_0' : \theta = 1/2$ versus $H_1 : \theta > 1/2$. ◄

The family of exponential distributions in Example 9.3.2 has a special property called *monotone likelihood ratio* that allows the manager to find a UMP test.

Definition 9.3.2 Monotone Likelihood Ratio. Let $f_n(\boldsymbol{x}|\theta)$ denote the joint p.d.f. or the joint p.f. of the observations $\boldsymbol{X} = (X_1, \ldots, X_n)$. Let $T = r(\boldsymbol{X})$ be a statistic. It is said that the joint distribution of $\boldsymbol{X}$ has a *monotone likelihood ratio (MLR) in the statistic T* if the following property is satisfied: For every two values $\theta_1 \in \Omega$ and $\theta_2 \in \Omega$, with $\theta_1 < \theta_2$, the ratio $f_n(\boldsymbol{x}|\theta_2)/f_n(\boldsymbol{x}|\theta_1)$ depends on the vector $\boldsymbol{x}$ only through the function $r(\boldsymbol{x})$,

and this ratio is a monotone function of $r(\boldsymbol{x})$ over the range of possible values of $r(\boldsymbol{x})$. Specifically, if the ratio is increasing, we say that the distribution of $\boldsymbol{X}$ has *increasing MLR*, and if the ratio is decreasing, we say that the distribution has *decreasing MLR*.

Example 9.3.3 **Sampling from a Bernoulli Distribution.** Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p $(0 < p < 1)$. If we let $y = \sum_{i=1}^{n} x_i$, then the joint p.f. $f_n(\boldsymbol{x}|p)$ is as follows:

$$f_n(\boldsymbol{x}|p) = p^y(1-p)^{n-y}.$$

Therefore, for every two values p_1 and p_2 such that $0 < p_1 < p_2 < 1$,

$$\frac{f_n(\boldsymbol{x}|p_2)}{f_n(\boldsymbol{x}|p_1)} = \left[\frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^y \left(\frac{1-p_2}{1-p_1}\right)^n. \tag{9.3.5}$$

It can be seen from Eq. (9.3.5) that the ratio $f_n(\boldsymbol{x}|p_2)/f_n(\boldsymbol{x}|p_1)$ depends on the vector $\boldsymbol{x}$ only through the value of y, and this ratio is an increasing function of y. Therefore, $f_n(\boldsymbol{x}|p)$ has increasing monotone likelihood ratio in the statistic $Y = \sum_{i=1}^{n} X_i$. ◄

Example 9.3.4 **Sampling from an Exponential Distribution.** Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the exponential distribution with unknown parameter $\theta > 0$. The joint p.d.f. is

$$f_n(\boldsymbol{x}|\theta) = \begin{cases} \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i\right) & \text{for all } x_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

For $0 < \theta_1 < \theta_2$, we have

$$\frac{f_n(\boldsymbol{x}|\theta_2)}{f_n(\boldsymbol{x}|\theta_1)} = \left(\frac{\theta_2}{\theta_1}\right)^n \exp\left([\theta_1 - \theta_2]\sum_{i=1}^{n} x_i\right), \tag{9.3.6}$$

if all $x_i > 0$. If we let $r(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$, then we see that the ratio in Eq. (9.3.6) depends on $\boldsymbol{x}$ only through $r(\boldsymbol{x})$ and is a decreasing function of $r(\boldsymbol{x})$. Hence, the joint distribution of a random sample of exponential random variables has decreasing MLR in $T = \sum_{i=1}^{n} X_i$. ◄

In Example 9.3.4, we could have defined the statistic $T' = -\sum_{i=1}^{n} X_i$ or $T' = 1/\sum_{i=1}^{n} X_i$, and then the distribution would have had increasing MLR in T'. This can be done in general in Definition 9.3.2. For this reason, when we prove theorems that assume that a distribution has MLR, we shall state and prove the theorems for increasing MLR only. When a distribution has decreasing MLR, the reader can transform the statistic by a strictly decreasing function and then transform the result back to the original statistic, if desired.

Example 9.3.5 **Sampling from a Normal Distribution.** Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ $(-\infty < \mu < \infty)$ and known variance σ^2. The joint p.d.f. $f_n(\boldsymbol{x}|\mu)$ is as follows:

$$f_n(\boldsymbol{x}|\mu) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right].$$

Therefore, for every two values μ_1 and μ_2 such that $\mu_1 < \mu_2$,

$$\frac{f_n(\boldsymbol{x}|\mu_2)}{f_n(\boldsymbol{x}|\mu_1)} = \exp\left\{\frac{n(\mu_2 - \mu_1)}{\sigma^2}\left[\overline{x}_n - \frac{1}{2}(\mu_2 + \mu_1)\right]\right\}. \tag{9.3.7}$$

It can be seen from Eq. (9.3.7) that the ratio $f_n(\boldsymbol{x}|\mu_2)/f_n(\boldsymbol{x}|\mu_1)$ depends on the vector $\boldsymbol{x}$ only through the value of $\bar{x}_n$, and this ratio is an increasing function of $\bar{x}_n$. Therefore, $f_n(\boldsymbol{x}|\mu)$ has increasing monotone likelihood ratio in the statistic $\overline{X}_n$. ◀

One-Sided Alternatives

In Example 9.3.2, we found a UMP level α_0 test for a simple null hypothesis $H_0' : \theta = 1/2$ against a one-sided alternative $H_1 : \theta > 1/2$. It is more common in such problems to test hypotheses of the form

$$\begin{aligned} H_0&: \quad \theta \le \theta_0, \\ H_1&: \quad \theta > \theta_0. \end{aligned} \tag{9.3.8}$$

That is, both the null and alternative hypotheses are one-sided. Because the one-sided null hypothesis is larger than the simple null $H_0' : \theta = \theta_0$, it is not necessarily the case that a level α_0 test of H_0' will be a level α_0 test of H_0. However, if the joint distribution of the observations has MLR, we will be able to show that there will exist UMP level α_0 tests of the hypotheses (9.3.8). Furthermore (see Exercise 12), there will exist UMP tests of the hypotheses obtained by reversing the inequalities in both H_0 and H_1 in (9.3.8).

Theorem 9.3.1 Suppose that the joint distribution of $\boldsymbol{X}$ has increasing monotone likelihood ratio in the statistic $T = r(\boldsymbol{X})$. Let c and α_0 be constants such that

$$\Pr(T \ge c|\theta = \theta_0) = \alpha_0. \tag{9.3.9}$$

Then the test procedure δ^* that rejects H_0 if $T \ge c$ is a UMP test of the hypotheses (9.3.8) at the level of significance α_0. Also, $\pi(\theta|\delta^*)$ is a monotone increasing function of θ.

Proof Let $\theta' < \theta''$ be arbitrary values of θ. Let $\alpha_0' = \pi(\theta'|\delta^*)$. It follows from the Nayman-Pearson lemma that among all procedures δ for which

$$\pi(\theta'|\delta) \le \alpha_0', \tag{9.3.10}$$

the value of $\pi(\theta''|\delta)$ will be maximized $(1 - \pi(\theta''|\delta)$ minimized) by a procedure that rejects H_0 when $f_n(\boldsymbol{x}|\theta'')/f_n(\boldsymbol{x}|\theta') \ge k$. The constant k is to be chosen so that

$$\pi(\theta'|\delta) = \alpha_0'. \tag{9.3.11}$$

Because the distribution of $\boldsymbol{X}$ has increasing MLR, the likelihood ratio $f_n(\boldsymbol{x}|\theta'')/f_n(\boldsymbol{x}|\theta')$ is an increasing function of $r(\boldsymbol{x})$. Therefore, a procedure that rejects H_0 when the likelihood ratio is at least equal to k will be equivalent to a procedure that rejects H_0 when $r(\boldsymbol{x})$ is at least equal to some other number c. The value of c is to be chosen so that (9.3.11) holds. The test δ^* satisfies Eq. (9.3.11) and has the correct form; hence, it maximizes the power function at $\theta = \theta''$ among all tests that satisfy Eq. (9.3.10). Another test δ that satisfies Eq. (9.3.10) is the following: Flip a coin that has probability of heads equal to α_0', and reject H_0 if the coin lands heads. This test has $\pi(\theta|\delta) = \alpha_0'$ for all θ including θ' and θ''. Because δ^* maximizes the power function at θ'', we have

$$\pi(\theta''|\delta^*) \ge \pi(\theta'|\delta) = \alpha_0' = \pi(\theta'|\delta^*). \tag{9.3.12}$$

Hence, we have proven the claim that $\pi(\theta|\delta^*)$ is a monotone increasing function of θ.

Next, consider the special case of what we have just proven with $\theta' = \theta_0$. Then $\alpha_0' = \alpha_0$, and we have proven that, for every $\theta'' > \theta_0$, δ^* maximizes $\pi(\theta''|\delta)$ among all

tests δ that satisfy

$$\pi(\theta_0|\delta) \le \alpha_0. \tag{9.3.13}$$

Every level α_0 test δ satisfies Eq. (9.3.13). Hence, δ^* has power at θ'' at least as high as the power of every level α_0 test. All that remains to complete the proof is to show that δ^* is itself a level α_0 test.

We have already shown that the power function $\pi(\theta|\delta^*)$ is monotone increasing. Hence, $\pi(\theta|\delta^*) \le \alpha_0$ for all $\theta \le \theta_0$, and δ^* is a level α_0 test. ■

Example 9.3.6 Service Times in a Queue. The manager in Example 9.3.2 might be interested in the hypotheses $H_0: \theta \le 1/2$ versus $H_1: \theta > 1/2$. The distribution in that example has decreasing MLR in the statistic $T = \sum_{i=1}^{n} X_i$, and hence it has increasing MLR in $-T$. Theorem 9.3.1 says that a UMP level α_0 test is to reject H_0 when $-T$ is greater than the $1 - \alpha_0$ quantile of the distribution of $-T$ given $\theta = 1/2$. This is the same as rejecting H_0 when T is less than the α_0 quantile of the distribution of T. The distribution of T given $\theta = 1/2$ is the gamma distribution with parameters n and 1/2, which is also the χ^2 distribution with $2n$ degrees of freedom. For example, if $n = 10$ and $\alpha_0 = 0.1$, the quantile is 12.44, which can be found in the table in the back of the book or from computer software. ◀

Example 9.3.7 Testing Hypotheses about the Proportion of Defective Items. Suppose that the proportion p of defective items in a large manufactured lot is unknown, 20 items are to be selected at random from the lot and inspected, and the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad p \le 0.1, \\ H_1&: \quad p > 0.1. \end{aligned} \tag{9.3.14}$$

We shall show first that there exist UMP tests of the hypotheses (9.3.14). We shall then determine the form of these tests and discuss the different levels of significance that can be attained with nonrandomized tests.

Let $X_1, \ldots, X_{20}$ denote the 20 random variables in the sample. Then $X_1, \ldots, X_{20}$ form a random sample of size 20 from the Bernoulli distribution with parameter p, and it is known from Example 9.3.3 that the joint p.f. of $X_1, \ldots, X_{20}$ has increasing monotone likelihood ratio in the statistic $Y = \sum_{i=1}^{20} X_i$. Therefore, by Theorem 9.3.1, a test procedure that rejects H_0 when $Y \ge c$ will be a UMP test of the hypotheses (9.3.14).

For each specific choice of the constant c, the size of the UMP test will be $\alpha_0 = \Pr(Y \ge c|p = 0.1)$. When $p = 0.1$, the random variable Y has the binomial distribution with parameters $n = 20$ and $p = 0.1$. Because Y has a discrete distribution and assumes only a finite number of different possible values, it follows that there are only a finite number of different possible values for α_0. To illustrate this remark, it is found from a table of the binomial distribution that if $c = 7$, then $\alpha_0 = \Pr(Y \ge 7|p = 0.1) = 0.0024$, and if $c = 6$, then $\alpha_0 = \Pr(Y \ge 6|p = 0.1) = 0.0113$. Therefore, if an experimenter wants the size of the test to be approximately 0.01, she could choose either $c = 7$ and $\alpha_0 = 0.0024$ or $c = 6$ and $\alpha_0 = 0.0113$. The test with $c = 7$ is a level 0.01 test while the test with $c = 6$ is not, because the size of the former test is less than 0.01 while the size of the latter test is greater than 0.01.

If the experimenter wants the size of the test to be exactly 0.01, then she can use a randomized test procedure of the type described in Sec. 9.2. ◀

Example 9.3.8 Testing Hypotheses about the Mean of a Normal Distribution. Let $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Assume

that σ^2 is known. Let μ_0 be a specified number, and suppose that the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad \mu \le \mu_0, \\ H_1&: \quad \mu > \mu_0. \end{aligned} \tag{9.3.15}$$

We shall show first that, for every specified level of significance α_0 $(0 < \alpha_0 < 1)$, there is a UMP test of the hypotheses (9.3.15) with size equal to α_0. We shall then determine the power function of the UMP test.

It is known from Example 9.3.5 that the joint p.d.f. of $X_1, \ldots, X_n$ has an increasing monotone likelihood ratio in the statistic $\overline{X}_n$. Therefore, by Theorem 9.3.1, a test procedure δ_1 that rejects H_0 when $\overline{X}_n \ge c$ is a UMP test of the hypotheses (9.3.15). The size of this test is $\alpha_0 = \Pr(\overline{X}_n \ge c|\mu = \mu_0)$.

Since $\overline{X}_n$ has a continuous distribution, c is the $1 - \alpha_0$ quantile of the distribution of $\overline{X}_n$ given $\mu = \mu_0$. That is, c is the $1 - \alpha_0$ quantile of the normal distribution with mean μ_0 and variance σ^2/n. As we learned in Chapter 5, this quantile is

$$c = \mu_0 + \Phi^{-1}(1 - \alpha_0)\sigma n^{-1/2}, \tag{9.3.16}$$

where Φ^{-1} is the quantile function of the standard normal distribution. For simplicity, we shall let $z_{\alpha_0} = \Phi^{-1}(1 - \alpha_0)$ for the rest of this example.

We shall now determine the power function $\pi(\mu|\delta_1)$ of this UMP test. By definition,

$$\pi(\mu|\delta_1) = \Pr(\text{Rejecting } H_0|\mu) = \Pr(\overline{X}_n \ge \mu_0 + z_{\alpha_0}\sigma n^{-1/2}|\mu). \tag{9.3.17}$$

For every value of μ, the random variable $Z' = n^{1/2}(\overline{X}_n - \mu)/\sigma$ will have the standard normal distribution. Therefore, if Φ denotes the c.d.f. of the standard normal distribution, then

$$\begin{aligned} \pi(\mu|\delta_1) &= \Pr\left[Z' \ge z_{\alpha_0} + \frac{n^{1/2}(\mu_0 - \mu)}{\sigma} \right] \\ &= 1 - \Phi\left[z_{\alpha_0} + \frac{n^{1/2}(\mu_0 - \mu)}{\sigma} \right] = \Phi\left[\frac{n^{1/2}(\mu - \mu_0)}{\sigma} - z_{\alpha_0} \right]. \end{aligned} \tag{9.3.18}$$

The power function $\pi(\mu|\delta_1)$ is sketched in Fig. 9.6. ◄

In each of the pairs of hypotheses (9.3.8), (9.3.14), and (9.3.15), the alternative hypothesis H_1 is called a *one-sided alternative* because the set of possible values of the parameter under H_1 lies entirely on one side of the set of possible values under the null hypothesis H_0. In particular, for the hypotheses (9.3.8), (9.3.14), or (9.3.15), every possible value of the parameter under H_1 is larger than every possible value under H_0.

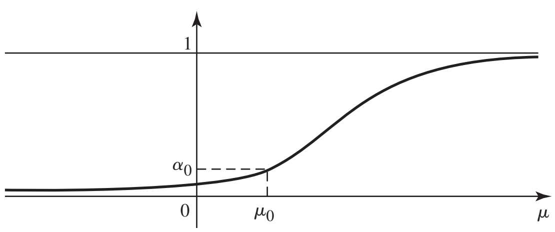


Figure 9.6 The power function $\pi(\mu|\delta_1)$ for the UMP test of the hypotheses (9.3.15).

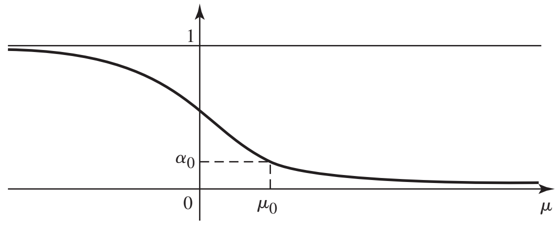


Figure 9.7 The power function $\pi(\mu|\delta_2)$ for the UMP test of the hypotheses (9.3.19).

Example 9.3.9 One-Sided Alternatives in the Other Direction. Suppose now that instead of testing the hypotheses (9.3.15) in Example 9.3.8, we are interested in testing the following hypotheses:

$$\begin{aligned} H_0&: \quad \mu \geq \mu_0, \\ H_1&: \quad \mu < \mu_0. \end{aligned} \tag{9.3.19}$$

In this case, the hypothesis H_1 is again a one-sided alternative, and it can be shown (see Exercise 12) that there exists a UMP test of the hypotheses (9.3.19) at every specified level of significance α_0 $(0 < \alpha_0 < 1)$. By analogy with Eq. (9.3.16), the UMP test δ_2 will reject H_0 when $\overline{X}_n \leq c$, where

$$c = \mu_0 - \Phi^{-1}(1 - \alpha_0)\sigma n^{-1/2}. \tag{9.3.20}$$

The power function $\pi(\mu|\delta_2)$ of the test δ_2 will be

$$\pi(\mu|\delta_2) = \Pr(\overline{X}_n \leq c|\mu) = \Phi\left[\frac{n^{1/2}(\mu_0 - \mu)}{\sigma} - \Phi^{-1}(1 - \alpha_0)\right]. \tag{9.3.21}$$

This function is sketched in Fig. 9.7. Indeed, Exercise 12 extends Theorem 9.3.1 to one-sided hypotheses of the form (9.3.19) in every monotone likelihood ratio family. In Sec. 9.8, we shall show that for all one-sided cases with monotone likelihood ratio, the tests of the form given in Theorem 9.3.1 and Exercise 12 are also optimal when one focuses on the posterior distribution of θ rather than on the power function. ◂

Two-Sided Alternatives

Suppose, finally, that instead of testing either the hypotheses (9.3.15) in Example 9.3.8 or the hypotheses (9.3.19), we are interested in testing the following hypotheses:

$$\begin{aligned} H_0&: \quad \mu = \mu_0, \\ H_1&: \quad \mu \neq \mu_0. \end{aligned} \tag{9.3.22}$$

In this case, H_0 is a simple hypothesis and H_1 is a two-sided alternative. Since H_0 is a simple hypothesis, the size of every test procedure δ will simply be equal to the value $\pi(\mu_0|\delta)$ of the power function at the point $\mu = \mu_0$.

Indeed, for each α_0 $(0 < \alpha_0 < 1)$, there is no UMP test of the hypotheses (9.3.22) at level of significance α_0. For every value of μ such that $\mu > \mu_0$, the value of $\pi(\mu|\delta)$ will be maximized by the test procedure δ_1 in Example 9.3.8, whereas for every value of μ such that $\mu < \mu_0$, the value of $\pi(\mu|\delta)$ will be maximized by the test procedure δ_2 in Example 9.3.9. It can be shown (see Exercise 19) that δ_1 is essentially the unique test that maximizes $\pi(\mu|\delta)$ for $\mu > \mu_0$. Since δ_1 does not maximize $\pi(\mu|\delta)$ for $\mu < \mu_0$, no test could maximize $\pi(\mu|\delta)$ simultaneously for $\mu > \mu_0$ and $\mu < \mu_0$. In the next section, we shall discuss the selection of an appropriate test procedure in this problem.

Summary

A uniformly most powerful (UMP) level α_0 test is a level α_0 test whose power function on the alternative hypothesis is always at least as high as the power function of every level α_0 test. If the family of distributions for the data has a monotone likelihood ratio in a statistic T, and if the null and alternative hypotheses are both one-sided, then there exists a UMP level α_0 test. In these cases, the UMP level α_0 test is either of the form "reject H_0 if $T \geq c$" or "reject H_0 if $T \leq c$."

Exercises

1. Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with unknown mean λ ($\lambda > 0$). Show that the joint p.f. of $X_1, \ldots, X_n$ has a monotone likelihood ratio in the statistic $\sum_{i=1}^n X_i$.

2. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with known mean μ and unknown variance σ^2 ($\sigma^2 > 0$). Show that the joint p.d.f. of $X_1, \ldots, X_n$ has a monotone likelihood ratio in the statistic $\sum_{i=1}^n (X_i - \mu)^2$.

3. Suppose that $X_1, \ldots, X_n$ form a random sample from the gamma distribution with parameters α and β. Assume that α is unknown ($\alpha > 0$) and that β is known. Show that the joint p.d.f. of $X_1, \ldots, X_n$ has a monotone likelihood ratio in the statistic $\prod_{i=1}^n X_i$.

4. Suppose that $X_1, \ldots, X_n$ form a random sample from the gamma distribution with parameters α and β. Assume that α is known and that β is unknown ($\beta > 0$). Show that the joint p.d.f. of $X_1, \ldots, X_n$ has a monotone likelihood ratio in the statistic $-\overline{X}_n$.

5. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution that belongs to an exponential family, as defined in Exercise 23 of Sec. 7.3, and the p.d.f. or the p.f. of this distribution is $f(\boldsymbol{x}|\theta)$, as given in that exercise. Suppose also that $c(\theta)$ is a strictly increasing function of θ. Show that the joint p.d.f. or the joint p.f. of $X_1, \ldots, X_n$ has a monotone likelihood ratio in the statistic $\sum_{i=1}^n d(X_i)$.

6. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$. Show that the joint p.d.f. of $X_1, \ldots, X_n$ has a monotone likelihood ratio in the statistic $\max\{X_1, \ldots, X_n\}$.

7. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution involving a parameter θ whose value is unknown, and suppose that it is desired to test the following hypotheses:

$$H_0:\quad \theta \leq \theta_0,$$
$$H_1:\quad \theta > \theta_0.$$

Suppose also that the test procedure to be used ignores the observed values in the sample and, instead, depends only on an auxiliary randomization in which an unbalanced coin is tossed so that a head will be obtained with probability 0.05, and a tail will be obtained with probability 0.95. If a head is obtained, then H_0 is rejected, and if a tail is obtained, then H_0 is not rejected. Describe the power function of this randomized test procedure.

8. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with known mean 0 and unknown variance σ^2, and suppose that it is desired to test the following hypotheses:

$$H_0:\quad \sigma^2 \leq 2,$$
$$H_1:\quad \sigma^2 > 2.$$

Show that there exists a UMP test of these hypotheses at every level of significance α_0 ($0 < \alpha_0 < 1$).

9. Show that the UMP test in Exercise 8 rejects H_0 when $\sum_{i=1}^n X_i^2 \geq c$, and determine the value of c when $n = 10$ and $\alpha_0 = 0.05$.

10. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p, and suppose that it is desired to test the following hypotheses:

$$H_0:\quad p \leq \tfrac{1}{2},$$
$$H_1:\quad p > \tfrac{1}{2}.$$

Show that if the sample size is $n = 20$, then there exists a nonrandomized UMP test of these hypotheses at the level of significance $\alpha_0 = 0.0577$ and at the level of significance $\alpha_0 = 0.0207$.

11. Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with unknown mean λ, and suppose that it is desired to test the following hypotheses:

$$H_0:\quad \lambda \leq 1,$$
$$H_1:\quad \lambda > 1.$$

Show that if the sample size is $n = 10$, then there exists a nonrandomized UMP test of these hypotheses at the level of significance $\alpha_0 = 0.0143$.

12. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution that involves a parameter θ whose value is unknown, and the joint p.d.f. or the joint p.f. $f_n(\boldsymbol{x}|\theta)$ has a monotone likelihood ratio in the statistic $T = r(\boldsymbol{X})$. Let θ_0

be a specified value of θ, and suppose that the following hypotheses are to be tested:

$$H_0:\quad \theta \ge \theta_0,$$
$$H_1:\quad \theta < \theta_0.$$

Let c be a constant such that $\Pr(T \le c|\theta = \theta_0) = \alpha_0$. Show that the test procedure which rejects H_0 if $T \le c$ is a UMP test at the level of significance α_0.

13. Suppose that four observations are taken at random from the normal distribution with unknown mean μ and known variance 1. Suppose also that the following hypotheses are to be tested:

$$H_0:\quad \mu \ge 10,$$
$$H_1:\quad \mu < 10.$$

- **a.** Determine a UMP test at the level of significance $\alpha_0 = 0.1$.
- **b.** Determine the power of this test when $\mu = 9$.
- **c.** Determine the probability of not rejecting H_0 if $\mu = 11$.

14. Suppose that $X_1, \dots, X_n$ form a random sample from the Poisson distribution with unknown mean λ, and suppose that it is desired to test the following hypotheses:

$$H_0:\quad \lambda \ge 1,$$
$$H_1:\quad \lambda < 1.$$

Suppose also that the sample size is $n = 10$. At what levels of significance α_0 in the interval $0 < \alpha_0 < 0.03$ do there exist nonrandomized UMP tests?

15. Suppose that $X_1, \dots, X_n$ form a random sample from the exponential distribution with unknown parameter β, and suppose that it is desired to test the following hypotheses:

$$H_0:\quad \beta \ge \tfrac{1}{2},$$
$$H_1:\quad \beta < \tfrac{1}{2}.$$

Show that at every level of significance α_0 $(0 < \alpha_0 < 1)$, there exists a UMP test that specifies rejecting H_0 when $\overline{X}_n \ge c$, for some constant c.

16. Consider again the conditions of Exercise 15, and suppose that the sample size is $n = 10$. Determine the value of the constant c that defines the UMP test at the level of significance $\alpha_0 = 0.05$. *Hint*: Use the table of the χ^2 distribution.

17. Consider a single observation X from the Cauchy distribution with unknown location parameter θ. That is, the p.d.f. of X is

$$f(x|\theta) = \frac{1}{\pi[1 + (x - \theta)^2]} \quad \text{for } -\infty < x < \infty.$$

Suppose that it is desired to test the following hypotheses:

$$H_0:\quad \theta = 0,$$
$$H_1:\quad \theta > 0.$$

Show that, for every α_0 $(0 < \alpha_0 < 1)$, there does not exist a UMP test of these hypotheses at level of significance α_0.

18. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance 1. Suppose also that the following hypotheses are to be tested:

$$H_0:\quad \mu \le 0,$$
$$H_1:\quad \mu > 0.$$

Let δ^* denote the UMP test of these hypotheses at the level of significance $\alpha_0 = 0.025$, and let $\pi(\mu|\delta^*)$ denote the power function of δ^*.

- **a.** Determine the smallest value of the sample size n for which $\pi(\mu|\delta^*) \ge 0.9$ for $\mu \ge 0.5$.
- **b.** Determine the smallest value of n for which $\pi(\mu|\delta^*) \le 0.001$ for $\mu \le -0.1$.

19. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance σ^2. In this problem, you will prove the missing steps from the proof that there is no UMP level α_0 test for the hypotheses in (9.3.22). Let δ_1 be the test procedure with level α_0 defined in Example 9.3.8.

- **a.** Let A be a set of possible values for the random vector $\boldsymbol{X} = (X_1 \dots, X_n)$. Let $\mu_1 \ne \mu_0$. Prove that $\Pr(\boldsymbol{X} \in A|\mu = \mu_0) > 0$ if and only if $\Pr(\boldsymbol{X} \in A|\mu = \mu_1) > 0$.
- **b.** Let δ be a size α_0 test for the hypotheses in (9.3.22) that differs from δ_1 in the following sense: There is a set A for which δ rejects its null hypothesis when $\boldsymbol{X} \in A$, δ_1 does not reject its null hypothesis when $\boldsymbol{X} \in A$, and $\Pr(\boldsymbol{X} \in A|\mu = \mu_0) > 0$. Prove that $\pi(\mu|\delta) < \pi(\mu|\delta_1)$ for all $\mu > \mu_0$.

★ 9.4 Two-Sided Alternatives

When testing a simple null hypothesis against a two-sided alternative (as at the end of Sec. 9.3), the choice of a test procedure requires a bit more care than in the one-sided case. This section discusses some of the issues and describes the most common choices.

General Form of the Procedure

Example 9.4.1

Egyptian Skulls. In Example 9.1.2, we considered how to compare measurements of skulls found in Egypt to modern measurements. For example, the average breadth of a modern-day skull is about 140mm. Suppose that we model the breadths of skulls from 4000 B.C. as normal random variables with unknown mean μ and known variance of 26. Unlike Example 9.1.6, suppose now that the researchers have no theory suggesting that skull breadths should increase over time. Instead, they are merely interested in whether breadths changed at all. How would they choose a test of the hypotheses $H_0 : \mu = 140$ versus $H_1 : \mu \neq 140$? ◂

In this section, we shall suppose that $\boldsymbol{X} = (X_1, \dots, X_n)$ is a random sample from a normal distribution for which the mean μ is unknown and the variance σ^2 is known, and that it is desired to test the following hypotheses:

$$\begin{aligned} H_0&: \quad \mu = \mu_0, \\ H_1&: \quad \mu \neq \mu_0. \end{aligned} \tag{9.4.1}$$

In most practical problems, we would assume that both μ and σ^2 were unknown. We shall address that case in Sec. 9.5.

It was claimed at the end of Sec. 9.3 that there is no UMP test of the hypotheses (9.4.1) at any specified level of significance α_0 $(0 < \alpha_0 < 1)$. Neither the test procedure δ_1 nor the procedure δ_2 defined in Examples 9.3.8 and 9.3.9 is appropriate for testing the hypotheses (9.4.1), because each of those procedures has high power function only on one side of two-sided alternative H_1 and they each have low power function on the other side. However, the properties of the procedures δ_1 and δ_2 given in Sec. 9.3 and the fact that the sample mean $\overline{X}_n$ is the M.L.E. of μ suggest that a reasonable test of the hypotheses (9.4.1) would be to reject H_0 if $\overline{X}_n$ is far from μ_0. In other words, it seems reasonable to use a test procedure δ that rejects H_0 if either $\overline{X}_n \le c_1$ or $\overline{X}_n \ge c_2$, where c_1 and c_2 are two suitably chosen constants, presumably with $c_1 < \mu_0$ and $c_2 > \mu_0$.

If the size of the test is to be α_0, then the values of c_1 and c_2 must be chosen so as to satisfy the following relation:

$$\Pr(\overline{X}_n \le c_1 | \mu = \mu_0) + \Pr(\overline{X}_n \ge c_2 | \mu = \mu_0) = \alpha_0. \tag{9.4.2}$$

There are an infinite number of pairs of values of c_1 and c_2 that satisfy Eq. (9.4.2). When $\mu = \mu_0$, the random variable $n^{1/2}(\overline{X}_n - \mu_0)/\sigma$ has the standard normal distribution. If, as usual, we let Φ denote the c.d.f. of the standard normal distribution, then it follows that Eq. (9.4.2) is equivalent to the following relation:

$$\Phi\left[\frac{n^{1/2}(c_1 - \mu_0)}{\sigma}\right] + 1 - \Phi\left[\frac{n^{1/2}(c_2 - \mu_0)}{\sigma}\right] = \alpha_0. \tag{9.4.3}$$

Corresponding to every pair of positive numbers α_1 and α_2 such that $\alpha_1 + \alpha_2 = \alpha_0$, there exists a pair of numbers c_1 and c_2 such that $\Phi[n^{1/2}(c_1 - \mu_0)/\sigma] = \alpha_1$ and $1 - \Phi[n^{1/2}(c_2 - \mu_0)/\sigma] = \alpha_2$. Every such pair of values of c_1 and c_2 will satisfy Eqs. (9.4.2) and (9.4.3).

For example, suppose that $\alpha_0 = 0.05$. Then, choosing $\alpha_1 = 0.025$ and $\alpha_2 = 0.025$ yields a test procedure δ_3, which is defined by the values $c_1 = \mu_0 - 1.96\sigma n^{-1/2}$ and $c_2 = \mu_0 + 1.96\sigma n^{-1/2}$. Also, choosing $\alpha_1 = 0.01$ and $\alpha_2 = 0.04$ yields a test procedure δ_4, which is defined by the values $c_1 = \mu_0 - 2.33\sigma n^{-1/2}$ and $c_2 = \mu_0 + 1.75\sigma n^{-1/2}$. The power functions $\pi(\mu|\delta_3)$ and $\pi(\mu|\delta_4)$ of these test procedures δ_3 and δ_4 are sketched

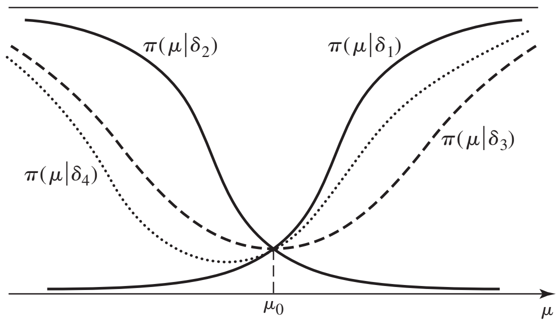


Figure 9.8 The power functions of four test procedures.

in Fig. 9.8, along with the power functions $\pi(\mu|\delta_1)$ and $\pi(\mu|\delta_2)$, which had previously been sketched in Figs. 9.6 and 9.7.

As the values of c_1 and c_2 in Eq. (9.4.2) or Eq. (9.4.3) are decreased, the power function $\pi(\mu|\delta)$ will become smaller for $\mu < \mu_0$ and larger for $\mu > \mu_0$. For $\alpha_0 = 0.05$, the limiting case is obtained by choosing $c_1 = -\infty$ and $c_2 = \mu_0 + 1.645\sigma n^{-1/2}$. The test procedure defined by these values is just δ_1. Similarly, as the values of c_1 and c_2 in Eq. (9.4.2) or Eq. (9.4.3) are increased, the power function $\pi(\mu|\delta)$ will become larger for $\mu < \mu_0$ and smaller for $\mu > \mu_0$. For $\alpha_0 = 0.05$, the limiting case is obtained by choosing $c_2 = \infty$ and $c_1 = \mu_0 - 1.645\sigma n^{-1/2}$. The test procedure defined by these values is just δ_2. Something between these two extreme limiting cases seems appropriate for hypotheses (9.4.1).

Selection of the Test Procedure

For a given sample size n, the values of the constants c_1 and c_2 in Eq. (9.4.2) should be chosen so that the size and shape of the power function are appropriate for the particular problem to be solved. In some problems, it is important not to reject the null hypothesis unless the data strongly indicate that μ differs greatly from μ_0. In such problems, a small value of α_0 should be used. In other problems, not rejecting the null hypothesis H_0 when μ is slightly larger than μ_0 is a more serious error than not rejecting H_0 when μ is slightly less than μ_0. Then it is better to select a test having a power function such as $\pi(\mu|\delta_4)$ in Fig. 9.8 than to select a test having a symmetric function such as $\pi(\mu|\delta_3)$.

In general, the choice of a particular test procedure in a given problem should be based both on the cost of rejecting H_0 when $\mu = \mu_0$ and on the cost, for each possible value of μ, of not rejecting H_0 when $\mu \neq \mu_0$. Also, when a test is being selected, the relative likelihoods of different values of μ should be considered. For example, if it is more likely that μ will be greater than μ_0 than that μ will be less than μ_0, then it is better to select a test for which the power function is large when $\mu > \mu_0$, and not so large when $\mu < \mu_0$, than to select one for which these relations are reversed.

Example 9.4.2

Egyptian Skulls. Suppose that, in Example 9.4.1, it is equally important to reject the null hypothesis that the mean breadth μ equals 140 when $\mu < 140$ as when $\mu > 140$. Then we should choose a test that rejects H_0 when the sample average $\overline{X}_n$ is either at most c_1 or at least c_2 where c_1 and c_2 are symmetric around 140. Suppose that we want a test of size $\alpha_0 = 0.05$. There are $n = 30$ skulls from 4000 B.C., so

$$c_1 = 140 - 1.96(26)^{1/2}30^{-1/2} = 138.18,$$

$$c_2 = 140 + 1.96(26)^{1/2}30^{-1/2} = 141.82.$$

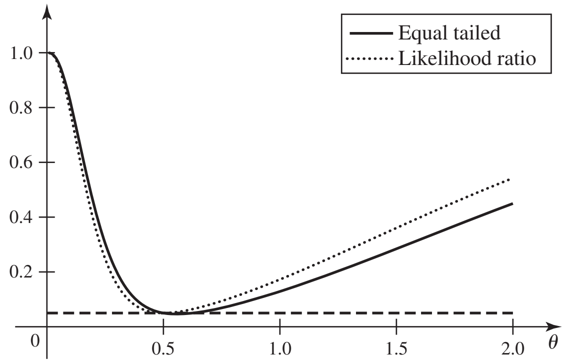


Figure 9.9 The power functions for the level $\alpha_0 = 0.05$ tests in Example 9.4.3 (equal tailed) and Example 9.4.4 (likelihood ratio). The horizontal line is at height 0.05.

The observed value of $\overline{X}_n$ is 131.37 in this case, and we would reject H_0 at the level of significance 0.05. ◄

In Examples 9.4.1 and 9.4.2, we would probably not wish to assume that the variance of the skull breadths was known to be 26, but rather we would assume that both the mean and the variance were unknown. We will see how to handle such a case in Sec. 9.5.

Other Distributions

The principles introduced above for samples from a normal distribution can be extended to any random sample. The details of implementation can be more tedious and less satisfying for other distributions.

Example 9.4.3 Service Times in a Queue. The manager in Example 9.3.2 models service times $X_1, \dots, X_n$ as i.i.d. exponential random variables with parameter θ conditional on θ. Suppose that she wishes to test the null hypothesis $H_0 : \theta = 1/2$ versus the alternative $H_1 : \theta \neq 1/2$. For the one-sided alternative $\theta > 1/2$, we found (in Example 9.3.2) that the UMP level α_0 test was to reject H_0 if $T = \sum_{i=1}^{n} X_i$ is less than the α_0 quantile of the gamma distribution with parameters n and 1/2. By similar reasoning, the UMP level α_0 test of H_0 versus the other one-sided alternative $\theta < 1/2$ would be to reject H_0 if T is greater than the $1 - \alpha_0$ quantile of the gamma distribution with parameters n and 1/2. A simple way to construct a level α_0 test of $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$ would be to apply the same reasoning that we applied immediately after Eq. (9.4.2). That is, combine two one-sided tests with levels α_1 and α_2 where $\alpha_1 + \alpha_2 = \alpha_0$.

As a specific example, let $\alpha_1 = \alpha_2 = \alpha_0/2$, and let $G^{-1}(\cdot; n, 1/2)$ be the quantile function of the gamma distribution with parameters n and 1/2. Then, we reject H_0 if $T \leq G^{-1}(\alpha_0/2; n, 1/2)$ or $T \geq G^{-1}(1 - \alpha_0/2; n, 1/2)$. For the case of $\alpha_0 = 0.05$ and $n = 3$, the graph of the power function of this test appears in Fig. 9.9 together with the power function of the likelihood ratio test that will be derived in Example 9.4.4. ◄

An alternative test in Example 9.4.3 would be the likelihood ratio test. In Example 9.4.3, the likelihood ratio test requires solving some nonlinear equations.

Example 9.4.4 Service Times in a Queue. Instead of the ad hoc two-sided test constructed in Example 9.4.3, suppose that the manager decides to find a likelihood ratio test. Suppose

that $\sum_{i=1}^{n} X_i = t$ is observed. The likelihood function is then

$$f_n(\boldsymbol{x}|\theta) = \theta^n \exp(-t\theta), \quad \text{for } \theta > 0.$$

The M.L.E. of θ is $\hat{\theta} = n/t$, so the likelihood ratio statistic from Definition 9.1.11 is

$$\Lambda(\boldsymbol{x}) = \frac{(1/2)^n \exp(-t/2)}{(n/t)^n \exp(-n)} = \left(\frac{t}{2n}\right)^n \exp(n - t/2). \tag{9.4.4}$$

The likelihood ratio test rejects H_0 if $\Lambda(\boldsymbol{x}) \le c$ for some constant c. From (9.4.4), we see that $\Lambda(\boldsymbol{x}) \le c$ is equivalent to $t \le c_1$ or $t \ge c_2$ where $c_1 < c_2$ satisfy

$$\left(\frac{c_1}{2n}\right)^n \exp(n - c_1/2) = \left(\frac{c_2}{2n}\right)^n \exp(n - c_2/2).$$

In order for the test to have level α_0, c_1 and c_2 must also satisfy

$$G(c_1; n, 1/2) + 1 - G(c_2; n, 1/2) = \alpha_0,$$

where $G(\cdot; n, 1/2)$ is the c.d.f. of the gamma distribution with parameters n and 1/2. Solving these two equations for c_1 and c_2 would give us the likelihood ratio test. Using numerical methods, the solution is $c_1 = 1.425$ and $c_2 = 15.897$. The power function of the likelihood ratio test is plotted in Fig. 9.9 together with the power function of the equal-tailed test. ◀

Composite Null Hypothesis

From one point of view, it makes little sense to carry out a test of the hypotheses (9.4.1) in which the null hypothesis H_0 specifies a single exact value μ_0 for the parameter μ. This is particularly true if we think of μ as the limit of the averages of increasing samples of future observations. Since it is inconceivable that μ will be *exactly* equal to μ_0 in any real problem, we know that the hypothesis H_0 cannot be true. Therefore, H_0 should be rejected as soon as it has been formulated.

This criticism is valid when it is interpreted literally. In many problems, however, the experimenter is interested in testing the null hypothesis H_0 that the value of μ is close to some specified value μ_0 against the alternative hypothesis that μ is not close to μ_0. In some of these problems, the simple hypothesis H_0 that $\mu = \mu_0$ can be used as an idealization or simplification for the purpose of choosing a decision. At other times, it is worthwhile to use a more realistic composite null hypothesis, which specifies that μ lies in an explicit interval around the value μ_0. We shall now consider hypotheses of this type.

Example 9.4.5

Testing an Interval Null Hypothesis. Suppose that $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance $\sigma^2 = 1$, and suppose that the following hypotheses are to be tested:

$$\begin{aligned} &H_0\colon\ 9.9 \le \mu \le 10.1,\\ &H_1\colon\ \mu < 9.9 \text{ or } \mu > 10.1. \end{aligned} \tag{9.4.5}$$

Since the alternative hypothesis H_1 is two-sided, it is again appropriate to use a test procedure δ that rejects H_0 if either $\overline{X}_n \le c_1$ or $\overline{X}_n \ge c_2$. We shall determine the values of c_1 and c_2 for which the probability of rejecting H_0, when either $\mu = 9.9$ or $\mu = 10.1$, will be 0.05.

Let $\pi(\mu|\delta)$ denote the power function of δ. When $\mu = 9.9$, the random variable $n^{1/2}(\overline{X}_n - 9.9)$ has the standard normal distribution. Therefore,

$$\begin{aligned}\pi(9.9|\delta) &= \Pr(\text{Rejecting } H_0|\mu = 9.9) \\ &= \Pr(\overline{X}_n \le c_1|\mu = 9.9) + \Pr(\overline{X}_n \ge c_2|\mu = 9.9) \\ &= \Phi[n^{1/2}(c_1 - 9.9)] + 1 - \Phi[n^{1/2}(c_2 - 9.9)].\end{aligned} \tag{9.4.6}$$

Similarly, when $\mu = 10.1$, the random variable $n^{1/2}(\overline{X}_n - 10.1)$ has the standard normal distribution and

$$\pi(10.1|\delta) = \Phi[n^{1/2}(c_1 - 10.1)] + 1 - \Phi[n^{1/2}(c_2 - 10.1)]. \tag{9.4.7}$$

Both $\pi(9.9|\delta)$ and $\pi(10.1|\delta)$ must be made equal to 0.05. Because of the symmetry of the normal distribution, it follows that if the values of c_1 and c_2 are chosen symmetrically with respect to the value 10, then the power function $\pi(\mu|\delta)$ will be symmetric with respect to the point $\mu = 10$. In particular, it will then be true that $\pi(9.9|\delta) = \pi(10.1|\delta)$.

Accordingly, let $c_1 = 10 - c$ and $c_2 = 10 + c$. Then it follows from Eqs. (9.4.6) and (9.4.7) that

$$\pi(9.9|\delta) = \pi(10.1|\delta) = \Phi[n^{1/2}(0.1 - c)] + 1 - \Phi[n^{1/2}(0.1 + c)]. \tag{9.4.8}$$

The value of c must be chosen so that $\pi(9.9|\delta) = \pi(10.1|\delta) = 0.05$. Therefore, c must be chosen so that

$$\Phi[n^{1/2}(0.1 + c)] - \Phi[n^{1/2}(0.1 - c)] = 0.95. \tag{9.4.9}$$

For each given value of n, the value of c that satisfies Eq. (9.4.9) can be found by trial and error from a table of the standard normal distribution or using statistical software.

For example, if $n = 16$, then c must be chosen so that

$$\Phi(0.4 + 4c) - \Phi(0.4 - 4c) = 0.95. \tag{9.4.10}$$

After trying various values of c, we find that Eq. (9.4.10) will be satisfied when $c = 0.527$. Hence,

$$c_1 = 10 - 0.527 = 9.473 \text{ and } c_2 = 10 + 0.527 = 10.527.$$

Thus, when $n = 16$, the procedure δ rejects H_0 when either $\overline{X}_n \le 9.437$ or $\overline{X}_n \ge 10.527$. This procedure has a power function $\pi(\mu|\delta)$, which is symmetric with respect to the point $\mu = 10$ and for which $\pi(9.9|\delta) = \pi(10.1|\delta) = 0.05$. Furthermore, it is true that $\pi(\mu|\delta) < 0.05$ for $9.9 < \mu < 10.1$ and $\pi(\mu|\delta) > 0.05$ for $\mu < 9.9$ or $\mu > 10.1$. The function $\pi(\mu|\delta)$ is sketched in Fig. 9.10. ◀

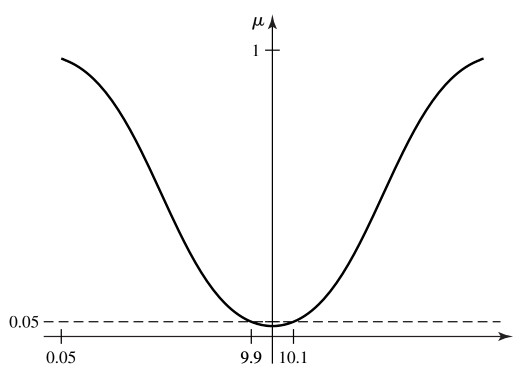


Figure 9.10 The power function $\pi(\mu|\delta)$ for a test of the hypotheses (9.4.5).

Unbiased Tests

Consider the general problem of testing the following hypotheses:

$$H_0: \quad \theta \in \Omega_0,$$
$$H_1: \quad \theta \in \Omega_1.$$

As usual, let $\pi(\theta|\delta)$ denote the power function of an arbitrary test procedure δ.

Definition 9.4.1 Unbiased Test. A test procedure δ is said to be *unbiased* if, for every $\theta \in \Omega_0$ and $\theta' \in \Omega_1$,

$$\pi(\theta|\delta) \le \pi(\theta'|\delta). \tag{9.4.11}$$

In words, δ is unbiased if its power function throughout Ω_1 is at least as large as it is throughout Ω_0.

If one closely examines Fig. 9.9, one sees that for values of θ slightly above 1/2, the power function of the equal-tailed test dips below 0.05 (the value of the power function at $\theta = 1/2$). This means that the test is not unbiased. This is typical in cases where the distribution of the test statistic T is not symmetric but a two-sided test is created by combining two one-sided tests. It is easy to see that an unbiased test would need to have a power function with derivative equal to 0 at $\theta = 1/2$; otherwise, it would dip below 0.05 on one side or the other of $\theta = 1/2$.

In many problems, the power function of every test is differentiable as a function of θ. In such cases, in order to create an unbiased level α_0 test δ of $H_0: \theta = \theta_0$ versus $H_1: \theta \ne \theta_0$, we would need

$$\pi(\theta_0|\delta) = \alpha_0, \text{ and}$$
$$\left.\frac{d}{d\theta}\pi(\theta|\delta)\right|_{\theta=\theta_0} = 0. \tag{9.4.12}$$

Such equations would need to be solved numerically in any real problem. Typically, researchers don't think it is worth the trouble to solve such equations just to find an unbiased test.

Example 9.4.6 Service Times in a Queue. In Example 9.4.4, let $T = \sum_{i=1}^n X_i$. If we want an unbiased test of the form "reject H_0 if $T \le c_1$ or if $T \ge c_2$," the power function will be

$$\pi(\theta|\delta) = G(c_1; n, \theta) + 1 - G(c_2; n, \theta),$$

where $G(\cdot; n, \theta)$ is the c.d.f. of T given θ,

$$G(x; n, \theta) = \int_0^x \frac{\theta^n}{(n-1)!} t^{n-1} \exp(-t\theta) dt,$$

for $t > 0$. Eq. (9.4.12) requires that we compute the derivative of G with respect to θ. The derivative with respect to θ can be passed under the integral, and the result is

$$\frac{\partial}{\partial\theta} G(x; n, \theta) = \int_0^x \frac{n\theta^{n-1}}{(n-1)!} t^{n-1} \exp(-t\theta) dt - \int_0^x t \frac{\theta^n}{(n-1)!} t^{n-1} \exp(-t\theta) dt. \tag{9.4.13}$$

The reader can show (see Exercise 13 in this section) that (9.4.13) can be rewritten as

$$\frac{\partial}{\partial\theta} G(x; n, \theta) = \frac{n}{\theta}\left[G(x; n, \theta) - G(x; n+1, \theta)\right]. \tag{9.4.14}$$

For $\alpha_0 = 0.05$ and $n = 3$, the two equations we need to solve for c_1 and c_2 are

$$G(c_1; 3, 1/2) + 1 - G(c_2; 3, 1/2) = 0.05,$$
$$\frac{3}{1/2}\left[G(x; 3, 1/2) - G(x; 4, 1/2)\right] = 0.$$

Solving these two equations numerically gives the same solution as the likelihood ratio test to the number of significant digits reported in Example 9.4.4. This explains why the power function of the likelihood ratio test appears not to dip below 0.05 anywhere. ◀

Intuitively, the notion of an unbiased test sounds appealing. Since the goal of a test procedure is to reject H_0 when $\theta \in \Omega_1$ and not to reject H_0 when $\theta \in \Omega_0$, it seems desirable that the probability of rejecting H_0 should be at least as large when $\theta \in \Omega_1$ as it is whenever $\theta \in \Omega_0$. It can be seen that the test δ for which the power function is sketched in Fig. 9.10 is an unbiased test of the hypotheses (9.4.5). Also, among the four tests for which the power functions are sketched in Fig. 9.8, only δ_3 is an unbiased test of the hypotheses (9.4.1). Although it is beyond the scope of this book, one can show that δ_3 is UMP among all unbiased level $\alpha_0 = 0.05$ tests of (9.4.1).

The requirement that a test is to be unbiased can sometimes narrow the selection of a test procedure. However, unbiased procedures should be sought only under relatively special circumstances. For example, when testing the hypotheses (9.4.5), the statistician should use the unbiased test δ represented in Fig. 9.10 only under the following conditions: He believes that, for every value $a > 0$, it is just as important to reject H_0 when $\theta = 10.1 + a$ as to reject H_0 when $\theta = 9.9 - a$, and he also believes that these two values of θ are equally likely. In practice, the statistician might very well forego the use of an unbiased test in order to use a biased test that has higher power in certain regions of Ω_1 that he regards as particularly important or most likely to contain the true value of θ when H_0 is false.

In the remainder of this chapter, we shall consider special testing situations that arise very often in applied work. In these situations, there do not exist UMP tests. We shall study the most popular tests in these situations, and we shall show that these tests are likelihood ratio tests. However, in more advanced courses, it can be shown that the t tests and F tests derived in Sections 9.5, 9.6, and 9.7 are all UMP among various classes of unbiased tests of their sizes.

Summary

For the case of testing that the mean of a normal distribution with known variance equals a specific value against the two-sided alternative, one can construct level α_0 tests by combining the rejection regions of two one-sided tests of sizes α_1 and α_2 such that $\alpha_0 = \alpha_1 + \alpha_2$. A popular choice is $\alpha_1 = \alpha_2 = \alpha_0/2$. In this case, if $X_1, \dots, X_n$ form a random sample from a normal distribution with mean μ and variance σ^2, one can test $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$ by rejecting H_0 if $\overline{X}_n > \mu_0 + \Phi^{-1}(1 - \alpha_0/2)\sigma/n^{1/2}$ or if $\overline{X}_n < \mu_0 - \Phi^{-1}(1 - \alpha_0/2)\sigma/n^{1/2}$, where Φ^{-1} is the quantile function of the standard normal distribution. A test is unbiased if its power function is greater at every point in the alternative hypothesis than at every point in the null hypothesis. The normal distribution test just described, with $\alpha_1 = \alpha_2 = \alpha_0/2$, is unbiased.

Exercises

1. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance 1, and it is desired to test the following hypotheses for a given number μ_0:

$$H_0: \quad \mu = \mu_0,$$
$$H_1: \quad \mu \neq \mu_0.$$

Consider a test procedure δ such that the hypothesis H_0 is rejected if either $\overline{X}_n \le c_1$ or $\overline{X}_n \ge c_2$, and let $\pi(\mu|\delta)$ denote the power function of δ. Determine the values of the constants c_1 and c_2 such that $\pi(\mu_0|\delta) = 0.10$ and the function $\pi(\mu|\delta)$ is symmetric with respect to the point $\mu = \mu_0$.

2. Consider again the conditions of Exercise 1, and suppose that

$$c_1 = \mu_0 - 1.96n^{-1/2}.$$

Determine the value of c_2 such that $\pi(\mu_0|\delta) = 0.10$.

3. Consider again the conditions of Exercise 1 and also the test procedure described in that exercise. Determine the smallest value of n for which $\pi(\mu_0|\delta) = 0.10$ and $\pi(\mu_0 + 1|\delta) = \pi(\mu_0 - 1|\delta) \ge 0.95$.

4. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance 1, and it is desired to test the following hypotheses:

$$H_0: \; 0.1 \le \mu \le 0.2,$$
$$H_1: \; \mu < 0.1 \text{ or } \mu > 0.2.$$

Consider a test procedure δ such that the hypothesis H_0 is rejected if either $\overline{X}_n \le c_1$ or $\overline{X}_n \ge c_2$, and let $\pi(\mu|\delta)$ denote the power function of δ. Suppose that the sample size is $n = 25$. Determine the values of the constants c_1 and c_2 such that $\pi(0.1|\delta) = \pi(0.2|\delta) = 0.07$.

5. Consider again the conditions of Exercise 4, and suppose also that $n = 25$. Determine the values of the constants c_1 and c_2 such that $\pi(0.1|\delta) = 0.02$ and $\pi(0.2|\delta) = 0.05$.

6. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown, and it is desired to test the following hypotheses:

$$H_0: \quad \theta \le 3,$$
$$H_1: \quad \theta > 3.$$

a. Show that for each level of significance α_0 $(0 \le \alpha_0 < 1)$, there exists a UMP test that specifies that H_0 should be rejected if $\max\{X_1, \ldots, X_n\} \ge c$.

b. Determine the value of c for each possible value of α_0.

7. For a given sample size n and a given value of α_0, sketch the power function of the UMP test found in Exercise 6.

8. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution described in Exercise 6, but suppose now that it is desired to test the following hypotheses:

$$H_0: \quad \theta \ge 3,$$
$$H_1: \quad \theta < 3.$$

a. Show that at each level of significance α_0 $(0 < \alpha_0 < 1)$, there exists a UMP test that specifies that H_0 should be rejected if $\max\{X_1, \ldots, X_n\} \le c$.

b. Determine the value of c for each possible value of α_0.

9. For a given sample size n and a given value of α_0, sketch the power function of the UMP test found in Exercise 8.

10. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution described in Exercise 6, but suppose now that it is desired to test the following hypotheses:

$$H_0: \quad \theta = 3, \qquad\qquad (9.4.15)$$
$$H_1: \quad \theta \neq 3.$$

Consider a test procedure δ such that the hypothesis H_0 is rejected if either $\max\{X_1, \ldots, X_n\} \le c_1$ or $\max\{X_1, \ldots, X_n\} \ge c_2$, and let $\pi(\theta|\delta)$ denote the power function of δ.

a. Determine the values of the constants c_1 and c_2 such that $\pi(3|\delta) = 0.05$ and δ is unbiased.

b. Prove that the test found in part (a) is UMP of level 0.05 for testing the hypotheses in (9.4.15). *Hint:* Compare this test to the UMP tests of level $\alpha_0 = 0.05$ in Exercises 6 and 8.

c. Determine the values of the constants c_1 and c_2 such that $\pi(3|\delta) = 0.05$ and δ is unbiased.

11. Consider again the conditions of Exercise 1. Determine the values of the constants c_1 and c_2 such that $\pi(\mu_0|\delta) = 0.10$ and δ is unbiased.

12. Let X have the exponential distribution with parameter β. Suppose that we wish to test the hypotheses

$$H_0: \quad \beta = 1,$$
$$H_1: \quad \beta \neq 1.$$

We shall use a test procedure that rejects H_0 if either $X \le c_1$ or $X \ge c_2$.

a. Find the equation that must be satisfied by c_1 and c_2 in order for the test procedure to have level of significance α_0.

b. Find a pair of finite, nonzero values (c_1, c_2) such that the test procedure has level of significance $\alpha_0 = 0.1$.

13. Prove Eq. (9.4.14) in Example 9.4.6.*Hint:* Both parts of the integrand in Eq. (9.4.13) differ from gamma distribution p.d.f.'s by some factor that does not depend on t.

9.5 The *t* Test

We begin the treatment of several special cases of testing hypotheses about parameters of a normal distribution. In this section, we handle the case in which both the mean and the variance are unknown. We develop tests for hypotheses concerning the mean. These tests will be based on the t distribution.

Testing Hypotheses about the Mean of a Normal Distribution When the Variance Is Unknown

Example 9.5.1 Nursing Homes in New Mexico. In Example 8.6.3, we described a study of medical in-patient days in nursing homes in New Mexico. As in that example, we shall model the numbers of medical in-patient days as a random sample of $n = 18$ normal random variables with unknown mean μ and unknown variance σ^2. Suppose that we are interested in testing the hypotheses $H_0 : \mu \geq 200$ versus $H_1 : \mu < 200$. What test should we use, and what are its properties? ◄

In this section we shall consider the problem of testing hypotheses about the mean of a normal distribution when both the mean and the variance are unknown. Specifically, we shall suppose that the random variables $X_1, \dots, X_n$ form a random sample from a normal distribution for which the mean μ and the variance σ^2 are unknown, and we shall consider testing the following hypotheses:

$$\begin{aligned} H_0&: \quad \mu \leq \mu_0, \\ H_1&: \quad \mu > \mu_0. \end{aligned} \tag{9.5.1}$$

The parameter space Ω in this problem comprises every two-dimensional vector (μ, σ^2), where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. The null hypothesis H_0 specifies that the vector (μ, σ^2) lies in the subset Ω_0 of Ω, comprising all vectors for which $\mu \leq \mu_0$ and $\sigma^2 > 0$, as illustrated in Fig. 9.11. The alternative hypothesis H_1 specifies that (μ, σ^2) belongs to the subset Ω_1 of Ω, comprising all the vectors that do not belong to Ω_0.

In Example 9.1.17 on page 543, we showed how to derive a test of the hypotheses (9.5.1) from a one-sided confidence interval for μ. To be specific, define $\overline{X}_n = \sum_{i=1}^n X_i/n$, $\sigma' = (\sum_{i=1}^n (X_i - \overline{X}_n)^2/[n-1])^{1/2}$, and

$$U = n^{1/2}\frac{\overline{X}_n - \mu_0}{\sigma'}. \tag{9.5.2}$$

The test rejects H_0 if $U \geq c$. When $\mu = \mu_0$, it follows from Theorem 8.4.2 that the distribution of the statistic U defined in Eq. (9.5.2) is the t distribution with $n - 1$

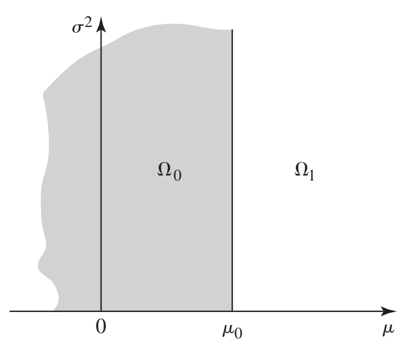


Figure 9.11 The subsets Ω_0 and Ω_1 of the parameter space Ω for the hypotheses (9.5.1).

degrees of freedom, regardless of the value of σ^2. For this reason, tests based on U are called *t tests*. When we want to test

$$\begin{aligned} H_0&: \quad \mu \ge \mu_0, \\ H_1&: \quad \mu < \mu_0, \end{aligned} \tag{9.5.3}$$

the test is of the form "reject H_0 if $U \le c$."

Example 9.5.2 Nursing Homes in New Mexico. In Example 9.5.1, if we desired a level α_0 test, we could use the t test that rejects H_0 if the statistic U in Eq. (9.5.2) is at most equal to the constant c chosen to make the size of the test equal to α_0. ◀

Properties of the t Tests

Theorem 9.5.1 gives some useful properties of t tests.

Theorem 9.5.1 Level and Unbiasedness of t Tests. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the normal distribution with mean μ and variance σ^2, let U be the statistic in Eq. (9.5.2), and let c be the $1 - \alpha_0$ quantile of the t distribution with $n - 1$ degrees of freedom. Let δ be the test that rejects H_0 in (9.5.1) if $U \ge c$. The power function $\pi(\mu, \sigma^2|\delta)$ has the following properties:

i. $\pi(\mu, \sigma^2|\delta) = \alpha_0$ when $\mu = \mu_0$,
ii. $\pi(\mu, \sigma^2|\delta) < \alpha_0$ when $\mu < \mu_0$,
iii. $\pi(\mu, \sigma^2|\delta) > \alpha_0$ when $\mu > \mu_0$,
iv. $\pi(\mu, \sigma^2|\delta) \to 0$ as $\mu \to -\infty$,
v. $\pi(\mu, \sigma^2|\delta) \to 1$ as $\mu \to \infty$.

Furthermore, the test δ has size α_0 and is unbiased.

Proof If $\mu = \mu_0$, then U has the t distribution with $n - 1$ degrees of freedom. Hence,

$$\pi(\mu_0, \sigma^2|\delta) = \Pr(U \ge c|\mu 0, \sigma^2) = \alpha_0.$$

This proves (i) above. For (ii) and (iii), define

$$U^* = \frac{n^{1/2}(\overline{X}_n - \mu)}{\sigma'} \quad \text{and} \quad W = \frac{n^{1/2}(\mu_0 - \mu)}{\sigma'}.$$

Then $U = U^* - W$. First, assume that $\mu < \mu_0$ so that $W > 0$. It follows that

$$\begin{aligned} \pi(\mu, \sigma^2|\delta) &= \Pr(U \ge c|\mu, \sigma^2) = \Pr(U^* - W \ge c|\mu, \sigma^2) \\ &= \Pr(U^* \ge c + W|\mu, \sigma^2) < \Pr(U^* \ge c|\mu, \sigma^2). \end{aligned} \tag{9.5.4}$$

Since U^* has the t distribution with $n - 1$ degrees of freedom, the last probability in (9.5.4) is α_0. This proves (ii). For (iii), let $\mu > \mu_0$ so that $W < 0$. The less-than in (9.5.4) becomes a greater-than, and (iii) is proven.

That the size of the test is α_0 is immediate from parts (i) and (ii). That the test is unbiased is immediate from parts (i) and (iii).

The proofs of (iv) and (v) are more difficult and will not be given here in detail. Intuitively, if μ is very large, then W in Eq. (9.5.4) will tend to be very negative, and the probability will be close to 1 that $U^* \ge c + W$. Similarly, if μ is very much less than 0, then W will tend to be very positive, and the chance of $U^* \ge c + W$ will be close to 0. ■

For the hypotheses of Eq. (9.5.3), very similar properties hold.

Corollary 9.5.1 t Tests for Hypotheses of Eq. (9.5.3). Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the normal distribution with mean μ and variance σ^2, let U be the statistic in Eq. (9.5.2), and let c be the α_0 quantile of the t distribution with $n-1$ degrees of freedom. Let δ be the test that rejects H_0 in (9.5.3) if $U \le c$. The power function $\pi(\mu, \sigma^2|\delta)$ has the following properties:

i. $\pi(\mu, \sigma^2|\delta) = \alpha_0$ when $\mu = \mu_0$,
ii. $\pi(\mu, \sigma^2|\delta) > \alpha_0$ when $\mu < \mu_0$,
iii. $\pi(\mu, \sigma^2|\delta) < \alpha_0$ when $\mu > \mu_0$,
iv. $\pi(\mu, \sigma^2|\delta) \to 1$ as $\mu \to -\infty$,
v. $\pi(\mu, \sigma^2|\delta) \to 0$ as $\mu \to \infty$.

Furthermore, the test δ has size α_0 and is unbiased.

Example 9.5.3 Nursing Homes in New Mexico. In Examples 9.5.1 and 9.5.2, suppose that we desire a test with level of significance $\alpha_0 = 0.1$. Then we reject H_0 if $U \le c$ where c is the 0.1 quantile of the t distribution with 17 degrees of freedom, namely, -1.333. Using the data from Example 8.6.3, we calculate the observed value of $\overline{X}_{18} = 182.17$ and $\sigma' = 72.22$. The observed value of U is then $(17)^{1/2}(182.17 - 200)/72.22 = -1.018$. We would not reject $H_0 : \mu \ge 200$ at level of significance 0.1, because the observed value of U is greater than -1.333. ◀

***p*-Values for *t* Tests** The p-value from the observed data and a specific test is the smallest α_0 such that we would reject the null hypothesis at level of significance α_0. For the t tests that we have just discussed, it is straightforward to compute the p-values.

Theorem 9.5.2 p-Values for t Tests. Suppose that we are testing either the hypotheses in Eq. (9.5.1) or the hypotheses in Eq. (9.5.3). Let u be the observed value of the statistic U in Eq. (9.5.2), and let $T_{n-1}(\cdot)$ be the c.d.f. of the t distribution with $n-1$ degrees of freedom. Then the p-value for the hypotheses in Eq. (9.5.1) is $1 - T_{n-1}(u)$ and the p-value for the hypotheses in Eq. (9.5.3) is $T_{n-1}(u)$.

Proof Let $T_{n-1}^{-1}(\cdot)$ stand for the quantile function of the t distribution with $n-1$ degrees of freedom. This is the inverse of the strictly increasing function T_{n-1}. We would reject the hypotheses in Eq. (9.5.1) at level α_0 if and only if $u \ge T_{n-1}^{-1}(1-\alpha_0)$, which is equivalent to $T_{n-1}(u) \ge 1 - \alpha_0$, which is equivalent to $\alpha_0 \ge 1 - T_{n-1}(u)$. Hence, the smallest level α_0 at which we could reject H_0 is $1 - T_{n-1}(u)$. Similarly, we would reject the hypotheses in Eq. (9.5.3) if and only if $u \le T_{n-1}^{-1}(\alpha_0)$, which is equivalent to $\alpha_0 \ge T_{n-1}(u)$. ■

Example 9.5.4 Lengths of Fibers. Suppose that the lengths in millimeters of metal fibers produced by a certain process have the normal distribution with unknown mean μ and unknown variance σ^2, and the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad \mu \le 5.2,\\ H_1&: \quad \mu > 5.2. \end{aligned} \tag{9.5.5}$$

Suppose that the lengths of 15 fibers selected at random are measured, and it is found that the sample mean $\overline{X}_{15}$ is 5.4 and $\sigma' = 0.4226$. Based on these measurements, we shall carry out a t test at the level of significance $\alpha_0 = 0.05$.

Since $n = 15$ and $\mu_0 = 5.2$, the statistic U defined by Eq. (9.5.2) will have the t distribution with 14 degrees of freedom when $\mu = 5.2$. It is found in the table of the

t distribution that $T_{14}^{-1}(0.95) = 1.761$. Hence, the null hypothesis H_0 will be rejected if $U > 1.761$. Since the numerical value of U calculated from Eq. (9.5.2) is 1.833, H_0 would be rejected at level 0.05.

With observed value $u = 1.833$ for the statistic U and $n = 15$, we can compute the p-value for the hypotheses (9.5.1) using computer software that includes the c.d.f. of various t distributions. In particular, we find $1 - T_{14}(1.833) = 0.0441$. ◀

The Complete Power Function For all values of μ, the power function of a t test can be determined if we know the distribution of U defined in Eq. (9.5.2). We can rewrite U as

$$U = \frac{n^{1/2}(\overline{X}_n - \mu_0)/\sigma}{\sigma'/\sigma}. \tag{9.5.6}$$

The numerator of the right side in Eq. (9.5.6) has the normal distribution with mean $n^{1/2}(\mu - \mu_0)/\sigma$ and variance 1. The denominator is the square-root of a χ^2 random variable divided by its degrees of freedom, $n - 1$. Were it not for the nonzero mean, the ratio would have the t distribution with $n - 1$ degrees of freedom as we have already shown. When the mean of the numerator is not 0, U has a *noncentral t distribution.*

Definition 9.5.1 Noncentral t Distributions. Let Y and W be independent random variables with W having the normal distribution with mean ψ and variance 1 and Y having the χ^2 distribution with m degrees of freedom. Then the distribution of

$$X = \frac{W}{\left(\frac{Y}{m}\right)^{1/2}},$$

is called the *noncentral t distribution with m degrees of freedom and noncentrality parameter* ψ. We shall let $T_m(t|\psi)$ denote the c.d.f. of this distribution. That is, $T_m(t|\psi) = \Pr(X \le t)$.

It should be obvious that the noncentral t distribution with m degrees of freedom and noncentrality parameter $\psi = 0$ is also the t distribution with m degrees of freedom. The following result is also immediate from Definition 9.5.1.

Theorem 9.5.3 Let $X_1, \dots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2. The distribution of the statistic U in Eq. (9.5.2) is the noncentral t distribution with $n - 1$ degrees of freedom and noncentrality parameter $\psi = n^{1/2}(\mu - \mu_0)/\sigma$. Let δ be the test that rejects $H_0 : \mu \le \mu_0$ when $U \ge c$. Then the power function of δ is $\pi(\mu, \sigma^2|\delta) = 1 - T_{n-1}(c|\psi)$. Let δ' be the test that rejects $H_0 : \mu \ge \mu_0$ when $U \le c$. Then the power function of δ' is $\pi(\mu, \sigma^2|\delta') = T_{n-1}(c|\psi)$. ■

In Exercise 11, you can prove that $1 - T_m(t|\psi) = T_m(-t| - \psi)$. There are computer programs to calculate the c.d.f.'s of noncentral t distributions, and some statistical software packages include such programs. Figure 9.12 plots the power functions of level 0.05 and level 0.01 t tests for various degrees of freedom and various values of the noncentrality parameter. The horizontal axis is labeled $|\psi|$ because the same graphs can be used for both types of one-sided hypotheses. The next example illustrates how to use Fig. 9.12 to approximate the power function.

Example 9.5.5 Lengths of Fibers. In Example 9.5.4, we tested the hypotheses (9.5.5) at level 0.05. Suppose that we are interested in the power of our test when μ is not equal to 5.2. In

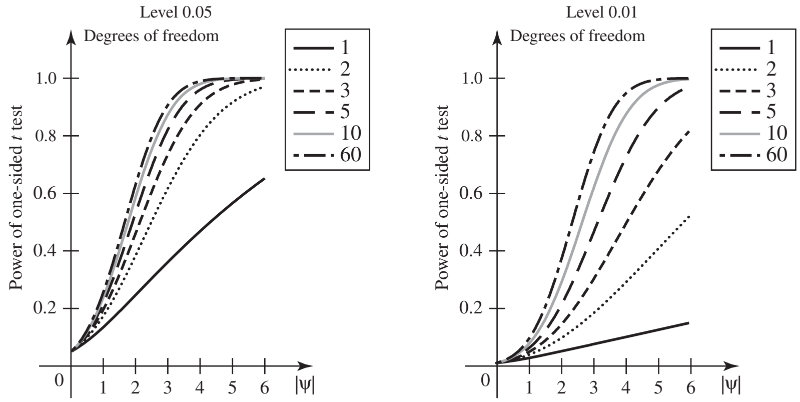


Figure 9.12 The power functions on the alternative of one-sided level 0.05 and level 0.01 t tests with various degrees of freedom for various values of the noncentrality parameter ψ.

particular, suppose that we are interested in the power when $\mu = 5.2 + \sigma/2$, one-half standard deviation above 5.2. Then the noncentrality parameter is

$$\psi = 15^{1/2}\left(\frac{5.2 + \sigma/2 - 5.2}{\sigma}\right) = 1.936.$$

There is no curve for 14 degrees of freedom in Fig. 9.12; however, there is not much difference between the curves for 10 and 60 degrees of freedom, so we can assume that our answer is somewhere between those two. If we look at the level 0.05 plot in Fig. 9.12 and move up from 1.936 (about 2) on the horizontal axis until we get a little above the curve for degrees of freedom equal to 10, we find that the power is about 0.6. (The actual power is 0.578.) ◄

Note: Power is a Function of the Noncentrality Parameter. In Example 9.5.5, we cannot answer a question like "What is the power of a level 0.05 test when $\mu = 5.5$?" The reason is that the power is a function of both μ and σ through the noncentrality parameter. (See Exercise 6.) For each possible σ and $\mu = 5.5$, the noncentrality parameter is $\psi = 15^{1/2} \times 0.3/\sigma$, which varies from 0 to ∞ depending on σ. This is why, whenever we want a numerical value for the power of a t test, we need either to specify both μ and σ or to specify how far μ is from μ_0 in multiples of σ.

Choosing a Sample Size It is possible to use the power function of a test to help determine what would be an appropriate sample size to observe.

Example 9.5.6

Lengths of Fibers. In Example 9.5.5, we found that the power of the test was 0.578 when $\mu = 5.2 + \sigma/2$. Suppose that we want the power to be close to 0.8, when $\mu = 5.2 + \sigma/2$. It will take more than $n = 15$ observations to achieve this. In Fig. 9.12, we can see what size of noncentrality parameter ψ that we need in order for the power to reach 0.8. For degrees of freedom between 10 and 60, we need ψ to be about 2.5. But $\psi = n^{1/2}/2$ when $\mu = 5.2 + \sigma/2$. So we need $n = 25$ approximately. Precise calculation shows that, with $n = 25$, the power of the level 0.05 test is 0.7834 when $\mu = 5.2 + \sigma/2$. With $n = 26$, the power is 0.7981, and with $n = 27$ the power is 0.8118. ◄

The Paired t Test

In many experiments, the same variable is measured under two different conditions on the same experimental unit, and we are interested in whether the mean value is

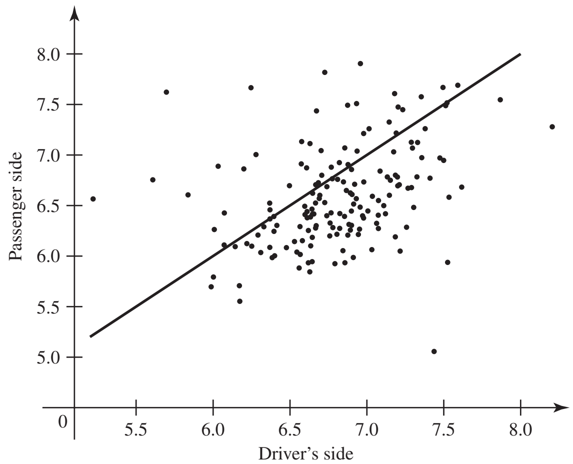


Figure 9.13 Plot of logarithms of head injury measures for dummies on driver's side and passenger's side. The line indicates where the two measures are equal.

greater in one condition than in the other. In such cases, it is common to subtract the two measurements and treat the differences as a random sample from a normal distribution. We can then test hypotheses concerning the mean of the differences.

Example 9.5.7

Crash Test Dummies. The National Transportation Safety Board collects data from crash tests concerning the amount and location of damage on dummies placed in the tested cars. In one series of tests, one dummy was placed in the driver's seat and another was placed in the front passenger's seat of each car. One variable measured was the amount of injury to the head for each dummy. Figure 9.13 shows a plot of the pairs of logarithms of head injury measures for dummies in the two different seats. Among other things, interest lies in whether and/or to what extent the amount of head injury differs between the driver's seat and the passenger's seat. Let $X_1, \dots, X_n$ be the differences between the logarithms of head injury measures for driver's side and passenger's side. We can model $X_1, \dots, X_n$ as a random sample from a normal distribution with mean μ and variance σ^2. Suppose that we wish to test the null hypothesis $H_0: \mu \le 0$ against the alternative $H_1: \mu > 0$ at level $\alpha_0 = 0.01$. There are $n = 164$ cars represented in Fig. 9.13. The test would be to reject H_0 if $U \ge T_{163}^{-1}(0.99) = 2.35$.

The average of the differences of the coordinates in Fig. 9.13 is $\overline{x}_n = 0.2199$. The value of σ' is 0.5342. The statistic U is then 5.271. This is larger than 2.35, and the null hypothesis would be rejected at level 0.01. Indeed, the p-value is less than 1.0×10^{-6}.

Suppose also that we are interested in the power function under H_1 of the level 0.01 test. Suppose that the mean difference between driver's side and passenger's side logarithm of head injury is $\sigma/4$. Then the noncentrality parameter is $(164)^{1/2}/4 = 3.20$. In the right panel of Fig. 9.12, it appears that the power is just about 0.8. (In fact, it is 0.802.) ◄

Testing with a Two-Sided Alternative

Example 9.5.8

Egyptian Skulls. In Examples 9.4.1 and 9.4.2, we modeled the breadths of skulls from 4000 B.C. as a random sample of size $n = 30$ from a normal distribution with unknown mean μ and known variance. We shall now generalize that model to allow the more realisitc assumption that the variance σ^2 is unknown. Suppose that we wish to test the null hypothesis $H_0: \mu = 140$ versus the alternative hypothesis $H_1: \mu \ne 140$. We can still calculate the statistic U in Eq. (9.5.2), but now it would make sense to reject

H_0 if either $U \le c_1$ or $U \ge c_2$ for suitably chosen numbers c_1 and c_2. How should we choose c_1 and c_2, and what are the properties of the resulting test? ◀

As before, assume that $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a random sample from a normal distribution for which both the mean μ and the variance σ^2 are unknown. Suppose now that the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad \mu = \mu_0, \\ H_1&: \quad \mu \ne \mu_0. \end{aligned} \tag{9.5.7}$$

Here, the alternative hypothesis H_1 is two-sided.

In Example 9.1.15, we derived a level α_0 test of the hypotheses (9.5.7) from the confidence interval that was developed in Sec. 8.5. That test has the form "reject H_0 if $|U| \ge T_{n-1}^{-1}(1 - \alpha_0/2)$," where T_{n-1}^{-1} is the quantile function of the t distribution with $n - 1$ degrees of freedom and U is defined in Eq. (9.5.2).

Example 9.5.9 Egyptian Skulls. In Example 9.5.8, suppose that we want a level $\alpha_0 = 0.05$ test of $H_0 : \mu = 140$ versus $H_1 : \mu \ne 140$. If we use the test described above (derived in Example 9.1.15), then the two numbers c_1 and c_2 will be of opposite signs and equal in magnitude. Specifically, $c_1 = -T_{29}^{-1}(0.975) = -2.045$ and $c_2 = 2.045$. The observed value of $\overline{X}_{30}$ is 131.37, and the observed value of σ' is 5.129. The observed value u of the statistic U is $u = (30)^{1/2}(131.37 - 140)/5.129 = -9.219$. This is less than -2.045, so we would reject H_0 at level 0.05. ◀

Example 9.5.10 Lengths of Fibers. We shall consider again the problem discussed in Example 9.5.4, but we shall suppose now that, instead of the hypotheses (9.5.5), the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad \mu = 5.2, \\ H_1&: \quad \mu \ne 5.2. \end{aligned} \tag{9.5.8}$$

We shall again assume that the lengths of 15 fibers are measured, and the value of U calculated from the observed values is 1.833. We shall test the hypotheses (9.5.8) at the level of significance $\alpha_0 = 0.05$.

Since $\alpha_0 = 0.05$, our critical value will be the $1 - 0.05/2 = 0.975$ quantile of the t distribution with 14 degrees of freedom. From the table of t distributions in this book, we find $T_{14}^{-1}(0.975) = 2.145$. So the t test specifies rejecting H_0 if either $U \le -2.145$ or $U \ge 2.145$. Since $U = 1.833$, the hypothesis H_0 would not be rejected. ◀

The numerical values in Examples 9.5.4 and 9.5.10 emphasize the importance of deciding whether the appropriate alternative hypothesis in a given problem is one-sided or two-sided. When the hypotheses (9.5.5) were tested at the level of significance 0.05, the hypothesis H_0 that $\mu \le 5.2$ was rejected. When the hypotheses (9.5.8) were tested at the same level of significance, and the same data were used, the hypothesis H_0 that $\mu = 5.2$ was not rejected.

Power Functions of Two-Sided Tests The power function of the test δ that rejects $H_0 : \mu = \mu_0$ when $|U| \ge c$, where $c = T_{n-1}^{-1}(1 - \alpha_0/2)$, can be found by using the noncentral t distribution. If $\mu \ne \mu_0$, then U has the noncentral t distribution with $n - 1$ degrees of freedom and noncentrality parameter $\psi = n^{1/2}(\mu - \mu_0)/\sigma$, just as it did when we tested one-sided hypotheses. The power function of δ is then

$$\pi(\mu, \sigma^2|\delta) = T_{n-1}(-c|\psi) + 1 - T_{n-1}(c|\psi).$$

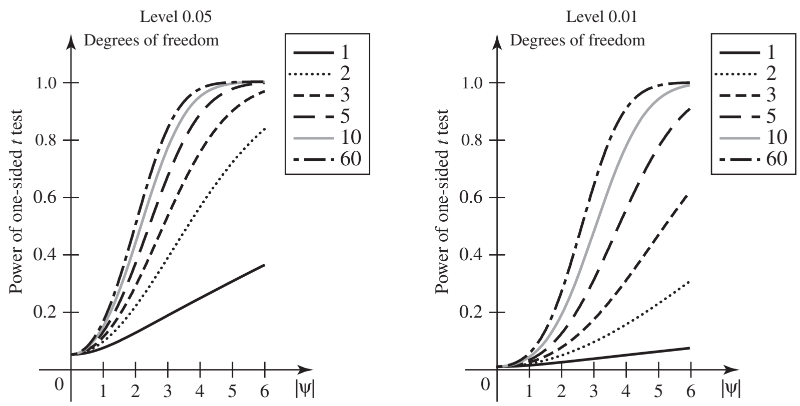


Figure 9.14 The power functions of two-sided level 0.05 and level 0.01 t tests with various degrees of freedom for various values of the noncentrality parameter ψ.

Figure 9.14 plots these power functions for various degrees of freedom and noncentrality parameters. We could use Fig. 9.14 to find the power of the test in Example 9.5.10 when $\mu = 5.2 + \sigma/2$, that is, when $\psi = 1.936$. It appears to be about 0.45. (The actual power is 0.438.)

Theorem 9.5.4 ***p*-Values for Two-Sided *t* Tests.** Suppose that we are testing the hypotheses in Eq. (9.5.7). Let u be the observed value of the statistic U, and let $T_{n-1}(\cdot)$ be the c.d.f. of the t distribution with $n-1$ degrees of freedom. Then the p-value is $2[1 - T_{n-1}(|u|)]$.

Proof Let $T_{n-1}^{-1}(\cdot)$ stand for the quantile function of the t distribution with $n-1$ degrees of freedom. We would reject the hypotheses in Eq. (9.5.7) at level α_0 if and only if $|u| \ge T_{n-1}^{-1}(1-\alpha_0/2)$, which is equivalent to $T_{n-1}(|u|) \ge 1-\alpha_0/2$, which is equivalent to $\alpha_0 \ge 2[1 - T_{n-1}(|u|)]$. Hence, the smallest level α_0 at which we could reject H_0 is $2[1 - T_{n-1}(|u|)]$. ■

Example 9.5.11 **Lengths of Fibers.** In Example 9.5.10, the p-value is $2[1 - T_{14}(1.833)] = 0.0882$. Note that this is twice the p-value when the hypotheses were (9.5.1). ◄

For t tests, if the p-value for testing hypotheses (9.5.1) or (9.5.3) is p, then the p-value for hypotheses (9.5.7) is the smaller of $2p$ and $2(1-p)$.

The *t* Test as a Likelihood Ratio Test

We introduced likelihood ratio tests in Sec. 9.1. We can compute such tests for the hypotheses of this section.

Example 9.5.12 **Likelihood Ratio Test of One-Sided Hypotheses about the Mean of a Normal Distribution.** Consider the hypotheses (9.5.1). After the values $x_1, \dots, x_n$ in the random sample have been observed, the likelihood function is

$$f_n(\boldsymbol{x}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]. \tag{9.5.9}$$

In this case, $\Omega_0 = \{(\mu, \sigma^2) : \mu \le \mu_0\}$ and $\Omega_1 = \{(\mu, \sigma^2) : \mu > \mu_0\}$. The likelihood ratio

statistic is

$$\Lambda(\boldsymbol{x}) = \frac{\sup_{\{(\mu,\sigma^2):\mu>\mu_0\}} f_n(\boldsymbol{x}|\mu, \sigma^2)}{\sup_{(\mu,\sigma^2)} f_n(\boldsymbol{x}|\mu, \sigma^2)}. \tag{9.5.10}$$

We shall now derive an explicit form for the likelihood ratio test based on (9.5.10). As in Sec. 7.5, we shall let $\hat{\mu}$ and $\hat{\sigma}^2$ denote the M.L.E.'s of μ and σ^2 when it is known only that the point (μ, σ^2) belongs to the parameter space Ω. It was shown in Example 7.5.6 that

$$\hat{\mu} = \overline{x}_n \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

It follows that the denominator of $\Lambda(\boldsymbol{x})$ equals

$$\sup_{(\mu,\sigma^2)} f_n(\boldsymbol{x}|\mu, \sigma^2) = \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}} \exp\left(-\frac{n}{2}\right). \tag{9.5.11}$$

Similarly, we shall let $\hat{\mu}_0$ and $\hat{\sigma}_0^2$ denote the M.L.E.'s of μ and σ^2 when the point (μ, σ^2) is constrained to lie in the subset Ω_0. Suppose first that the observed sample values are such that $\overline{x}_n \le \mu_0$. Then the point $(\hat{\mu}, \hat{\sigma}^2)$ will lie in Ω_0 so that $\hat{\mu}_0 = \hat{\mu}$ and $\hat{\sigma}_0^2 = \hat{\sigma}^2$ and the numerator of $\Lambda(\boldsymbol{x})$ also equals (9.5.11). In this case, $\Lambda(\boldsymbol{x}) = 1$.

Next, suppose that the observed sample values are such that $\overline{x}_n > \mu_0$. Then the point $(\hat{\mu}, \hat{\sigma}^2)$ does not lie in Ω_0. In this case, it can be shown that $f_n(\boldsymbol{x}|\mu, \sigma^2)$ attains its maximum value among all points $(\mu, \sigma^2) \in \Omega_0$ if μ is chosen to be as close as possible to $\overline{x}_n$. The value of μ closest to $\overline{x}_n$ among all points in the subset Ω_0 is $\mu = \mu_0$. Hence, $\hat{\mu}_0 = \mu_0$. In turn, it can be shown, as in Example 7.5.6, that the M.L.E. of σ^2 will be

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_0)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2.$$

In this case, the numerator of $\Lambda(\boldsymbol{x})$ is then

$$\sup_{\{(\mu,\sigma^2):\mu>\mu_0\}} f_n(\boldsymbol{x}|\mu, \sigma^2) = \frac{1}{(2\pi\hat{\sigma}_0^2)^{n/2}} \exp\left(-\frac{n}{2}\right). \tag{9.5.12}$$

Taking the ratio of (9.5.12) to (9.5.11), we find that

$$\Lambda(\boldsymbol{x}) = \begin{cases} \left(\dfrac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} & \text{if } \overline{x}_n > \mu_0, \\ 1 & \text{otherwise.} \end{cases} \tag{9.5.13}$$

Next, use the relation

$$\sum_{i=1}^{n}(x_i - \mu_0)^2 = \sum_{i=1}^{n}(x_i - \overline{x}_n)^2 + n(\overline{x}_n - \mu_0)^2$$

to write the top branch of (9.5.13) as

$$\left[1 + \frac{n(\overline{x}_n - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}\right]^{-n/2}. \tag{9.5.14}$$

If u is the observed value of the statistic U in Eq. (9.5.2), then one can easily check that

$$\frac{n(\overline{x}_n - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2} = \frac{u^2}{n-1}.$$

It follows that $\Lambda(\boldsymbol{x})$ is a nonincreasing function of u. Hence, for $k < 1$, $\Lambda(\boldsymbol{x}) \le k$ if and only if $u \ge c$, where

$$c = \left(\left[\frac{1}{k^{2/n}} - 1\right](n-1)\right)^{1/2}.$$

It follows that the likelihood ratio test is a t test. ◀

It is not difficult to adapt the argument in Example 9.5.12 to find the likelihood ratio tests for hypotheses (9.5.3) and (9.5.7). (See Exercises 17 and 18, for example.)

Summary

When $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2, we can test hypotheses about μ by using the fact that $n^{1/2}(\overline{X}_n - \mu)/\sigma'$ has the t distribution with $n-1$ degrees of freedom. Let T_{n-1}^{-1} denote the quantile function of the t distribution with $n-1$ degrees of freedom. Then, to test $H_0: \mu \le \mu_0$ versus $H_1: \mu > 0$ at level α_0, for instance, we reject H_0 if $n^{1/2}(\overline{X}_n - \mu_0)/\sigma' > T_{n-1}^{-1}(1-\alpha_0)$. To test $H_0: \mu = \mu_0$ versus $H_1: \mu \ne \mu_0$, reject H_0 if $|n^{1/2}(\overline{X}_n - \mu_0)/\sigma'| \ge T_{n-1}^{-1}(1-\alpha_0/2)$. The power functions of each of these tests can be written in terms of the c.d.f. of a noncentral t distribution with $n-1$ degrees of freedom and noncentrality parameter $\psi = n^{1/2}(\mu - \mu_0)/\sigma$.

Exercises

1. Use the data in Example 8.5.4, comprising a sample of $n = 10$ lactic acid measurements in cheese. Assume, as we did there, that the lactic acid measurements are a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Suppose that we wish to test the following hypotheses:

$$H_0: \quad \mu \le 1.2,$$
$$H_1: \quad \mu > 1.2.$$

a. Perform the level $\alpha_0 = 0.05$ test of these hypotheses.

b. Compute the p-value.

2. Suppose that nine observations are selected at random from the normal distribution with unknown mean μ and unknown variance σ^2, and for these nine observations it is found that $\overline{X}_n = 22$ and $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 72$.

a. Carry out a test of the following hypotheses at the level of significance 0.05:

$$H_0: \quad \mu \le 20,$$
$$H_1: \quad \mu > 20.$$

b. Carry out a test of the following hypotheses at the level of significance 0.05 by using the two-sided t test:

$$H_0: \quad \mu = 20,$$
$$H_1: \quad \mu \ne 20.$$

c. From the data, construct the observed confidence interval for μ with confidence coefficient 0.95.

3. The manufacturer of a certain type of automobile claims that under typical urban driving conditions the automobile will travel on average at least 20 miles per gallon of gasoline. The owner of this type of automobile notes the mileages that she has obtained in her own urban driving when she fills her automobile's tank with gasoline on nine different occasions. She finds that the results, in miles per gallon, are as follows: 15.6, 18.6, 18.3, 20.1, 21.5, 18.4, 19.1, 20.4, and 19.0. Test the manufacturer's claim by carrying out a test at the level of significance $\alpha_0 = 0.05$. List carefully the assumptions you make.

4. Suppose that a random sample of eight observations $X_1, \dots, X_8$ is taken from the normal distribution with unknown mean μ and unknown variance σ^2, and it is desired to test the following hypotheses:

$$H_0: \quad \mu = 0,$$
$$H_1: \quad \mu \ne 0.$$

Suppose also that the sample data are such that $\sum_{i=1}^{8} X_i = -11.2$ and $\sum_{i=1}^{8} X_i^2 = 43.7$. If a symmetric t test is performed at the level of significance 0.10 so that each tail of the critical region has probability 0.05, should the hypothesis H_0 be rejected or not?

5. Consider again the conditions of Exercise 4, and suppose again that a t test is to be performed at the level of significance 0.10. Suppose now, however, that the t test is not to be symmetric and the hypothesis H_0 is to be rejected if either $U \le c_1$ or $U \ge c_2$, where $\Pr(U \le c_1) = 0.01$ and $\Pr(U \ge c_2) = 0.09$. For the sample data specified in Exercise 4, should H_0 be rejected or not?

6. Suppose that the variables $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2, and a t test at a given level of significance α_0 is to be carried out to test the following hypotheses:

$$\begin{aligned} H_0&: \quad \mu \le \mu_0,\\ H_1&: \quad \mu > \mu_0. \end{aligned}$$

Let $\pi(\mu, \sigma^2|\delta)$ denote the power function of this t test, and assume that (μ_1, σ_1^2) and (μ_2, σ_2^2) are values of the parameters such that

$$\frac{\mu_1 - \mu_0}{\sigma_1} = \frac{\mu_2 - \mu_0}{\sigma_2}.$$

Show that $\pi(\mu_1, \sigma_1^2|\delta) = \pi(\mu_2, \sigma_2^2|\delta)$.

7. Consider the normal distribution with unknown mean μ and unknown variance σ^2, and suppose that it is desired to test the following hypotheses:

$$\begin{aligned} H_0&: \quad \mu \le \mu_0,\\ H_1&: \quad \mu > \mu_0. \end{aligned}$$

Suppose that it is possible to observe only a single value of X from this distribution, but that an independent random sample of n observations $Y_1, \dots, Y_n$ is available from the normal distribution with known mean 0 and the same variance σ^2 as for X. Show how to carry out a test of the hypotheses H_0 and H_1 based on the t distribution with n degrees of freedom.

8. Suppose that the variables $X_1, \dots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Let σ_0^2 be a given positive number, and suppose that it is desired to test the following hypotheses at a specified level of significance α_0 $(0 < \alpha_0 < 1)$:

$$\begin{aligned} H_0&: \quad \sigma^2 \le \sigma_0^2,\\ H_1&: \quad \sigma^2 > \sigma_0^2. \end{aligned}$$

Let $S_n^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2$, and suppose that the test procedure to be used specifies that H_0 should be rejected if $S_n^2/\sigma_0^2 \ge c$. Also, let $\pi(\mu, \sigma^2|\delta)$ denote the power function of this procedure. Explain how to choose the constant c so that, regardless of the value of μ, the following requirements are satisfied: $\pi(\mu, \sigma^2|\delta) < \alpha_0$ if $\sigma^2 < \sigma_0^2$, $\pi(\mu, \sigma^2|\delta) = \alpha_0$ if $\sigma^2 = \sigma_0^2$, and $\pi(\mu, \sigma^2|\delta) > \alpha_0$ if $\sigma^2 > \sigma_0^2$.

9. Suppose that a random sample of 10 observations $X_1, \dots, X_{10}$ is taken from the normal distribution with unknown mean μ and unknown variance σ^2, and it is desired to test the following hypotheses:

$$\begin{aligned} H_0&: \quad \sigma^2 \le 4,\\ H_1&: \quad \sigma^2 > 4. \end{aligned}$$

Suppose that a test of the form described in Exercise 8 is to be carried out at the level of significance $\alpha_0 = 0.05$. If the observed value of S_n^2 is 60, should the hypothesis H_0 be rejected or not?

10. Suppose again, as in Exercise 9, that a random sample of 10 observations is taken from the normal distribution with unknown mean μ and unknown variance σ^2, but suppose now that the following hypotheses are to be tested at the level of significance 0.05:

$$\begin{aligned} H_0&: \quad \sigma^2 = 4,\\ H_1&: \quad \sigma^2 \ne 4. \end{aligned}$$

Suppose that the null hypothesis H_0 is to be rejected if either $S_n^2 \le c_1$ or $S_n^2 \ge c_2$, where the constants c_1 and c_2 are to be chosen so that, when the hypothesis H_0 is true,

$$\Pr(S_n^2 \le c_1) = \Pr(S_n^2 \ge c_2) = 0.025.$$

Determine the values of c_1 and c_2.

11. Suppose that U_1 has the noncentral t distribution with m degrees of freedom and noncentrality parameter ψ, and suppose that U_2 has the noncentral t distribution with m degrees of freedom and noncentrality parameter $-\psi$. Prove that $\Pr(U_1 \ge c) = \Pr(U_2 \le -c)$.

12. Suppose that a random sample $X_1, \dots, X_n$ is to be taken from the normal distribution with unknown mean μ and unknown variance σ^2, and the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad \mu \le 3,\\ H_1&: \quad \mu > 3. \end{aligned}$$

Suppose also that the sample size n is 17, and it is found from the observed values in the sample that $\overline{X}_n = 3.2$ and $(1/n)\sum_{i=1}^n (X_i - \overline{X}_n)^2 = 0.09$. Calculate the value of the statistic U, and find the corresponding p-value.

13. Consider again the conditions of Exercise 12, but suppose now that the sample size n is 170, and it is again found from the observed values in the sample that $\overline{X}_n = 3.2$ and $(1/n)\sum_{i=1}^n (X_i - \overline{X}_n)^2 = 0.09$. Calculate the value of the statistic U and find the corresponding p-value.

14. Consider again the conditions of Exercise 12, but suppose now that the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad \mu = 3.1,\\ H_1&: \quad \mu \ne 3.1. \end{aligned}$$

Suppose, as in Exercise 12, that the sample size n is 17, and it is found from the observed values in the sample that

$\overline{X}_n = 3.2$ and $(1/n)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 0.09$. Calculate the value of the statistic U and find the corresponding p-value.

15. Consider again the conditions of Exercise 14, but suppose now that the sample size n is 170, and it is again found from the observed values in the sample that $\overline{X}_n = 3.2$ and $(1/n)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 0.09$. Calculate the value of the statistic U and find the corresponding p-value.

16. Consider again the conditions of Exercise 14. Suppose, as in Exercise 14, that the sample size n is 17, but suppose now that it is found from the observed values in the sample that $\overline{X}_n = 3.0$ and $(1/n)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 0.09$. Calculate the value of the statistic U and find corresponding p-value.

17. Prove that the likelihood ratio test for hypotheses (9.5.7) is the two-sided t test that rejects H_0 if $|U| \geq c$, where U is defined in Eq. (8.5.1). The argument is slightly simpler than, but very similar to, the one given in the text for the one-sided case.

18. Prove that the likelihood ratio test for hypotheses (9.5.3) is to reject H_0 if $U \leq c$, where U is defined in Eq. (8.5.1).

9.6 Comparing the Means of Two Normal Distributions

It is very common to compare two distributions to see which has the higher mean or just to see how different the two means are. When the two distributions are normal, the tests and confidence intervals based on the t distribution are very similar to the ones that arose when we considered a single distribution.

The Two-Sample t Test

Example 9.6.1 Rain from Seeded Clouds. In Example 8.3.1, we were interested in whether or not the mean log-rainfall from seeded clouds was greater than 4, which we supposed to have been the mean log-rainfall from unseeded clouds. If we want to compare rainfalls from seeded and unseeded clouds under otherwise similar conditions, we would normally observe two random samples of rainfalls: one from seeded clouds and one from unseeded clouds but otherwise under similar conditions. We would then model these samples as being random samples from two different normal distributions, and we would want to compare their means and possibly their variances to see how different the distributions are. ◀

Consider first a problem in which random samples are available from two normal distributions with common unknown variance, and it is desired to determine which distribution has the larger mean. Specifically, we shall assume that $\boldsymbol{X} = (X_1, \ldots, X_m)$ form a random sample of m observations from a normal distribution for which both the mean μ_1 and the variance σ^2 are unknown, and that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ form an independent random sample of n observations from another normal distribution for which both the mean μ_2 and the variance σ^2 are unknown. We will then be interested in testing hypotheses such as

$$H_0\colon \mu_1 \leq \mu_2 \quad \text{versus} \quad H_1\colon \mu_1 > \mu_2. \tag{9.6.1}$$

For each test procedure δ, we shall let $\pi(\mu_1, \mu_2, \sigma^2|\delta)$ denote the power function of δ. We shall assume that the variance σ^2 is the same for both distributions, even though the value of σ^2 is unknown. If this assumption seems unwarranted, the two-sample t test that we shall derive next would not be appropriate. A different test procedure is discussed later in this section for the case in which the two populations might have different variances. Later in this section, we shall derive the likelihood ratio test. In Sec. 9.7, we discuss some procedures for comparing the variances of two normal

distributions, which includes testing the null hypothesis that the variances are the same.

Intuitively, it makes sense to reject H_0 in (9.6.1) if the difference between the sample means is large. Theorem 9.6.1 derives the distribution of a natural test statistic to use.

Theorem 9.6.1 Two-Sample t Statistic. Assume the structure described in the preceding paragraphs. Define

$$\overline{X}_m = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$S_X^2 = \sum_{i=1}^{m}(X_i - \overline{X}_m)^2, \quad \text{and} \quad S_Y^2 = \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2. \tag{9.6.2}$$

Define the test statistic

$$U = \frac{(m+n-2)^{1/2}(\overline{X}_m - \overline{Y}_n)}{\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}(S_X^2 + S_Y^2)^{1/2}}. \tag{9.6.3}$$

For all values of $\theta = (\mu_1, \mu_2, \sigma^2)$ such that $\mu_1 = \mu_2$, the distribution of U is the t distribution with $m + n - 2$ degrees of freedom.

Proof Assume that $\mu_1 = \mu_2$. Define the following two random variables:

$$Z = \frac{\overline{X}_m - \overline{Y}_n}{\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}\sigma}, \tag{9.6.4}$$

$$W = \frac{S_X^2 + S_Y^2}{\sigma^2}. \tag{9.6.5}$$

The statistic U can now be represented in the form

$$U = \frac{Z}{[W/(m+n-2)]^{1/2}}. \tag{9.6.6}$$

The remainder of the proof consists of proving that Z has the standard normal distribution, that W has the χ^2 distribution with $m + n - 2$ degrees of freedom, and that Z and W are independent. The result then follows from Definition 8.4.1, the definition of the family of t distributions.

We have assumed that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent given θ. It follows that every function of $\boldsymbol{X}$ is independent of every function of $\boldsymbol{Y}$. In particular, $(\overline{X}_m, S_X^2)$ is independent of $(\overline{Y}_n, S_Y^2)$. By Theorem 8.3.1, $\overline{X}_m$ and S_X^2 are independent, and $\overline{Y}_n$ and S_Y^2 are also independent. It follows that all four of $\overline{X}_m$, $\overline{Y}_n$, S_X^2, and S_Y^2 are mutually independent. Hence, Z and W are also independent. It also follows from Theorem 8.3.1 that S_X^2/σ^2 and S_Y^2/σ^2 have, respectively, the χ^2 distributions with $m-1$ and $n-1$ degrees of freedom. Hence, W is the sum of two independent random variables with χ^2 distributions and so has the χ^2 distribution with the sum of the two degrees of freedom, namely, $m + n - 2$. $\overline{X}_m - \overline{Y}_n$ has the normal distribution with mean $\mu_1 - \mu_2 = 0$ and variance $\sigma^2/n + \sigma^2/m$. It follows that Z has the standard normal distribution. ■

A two-sample t test with level of significance α_0 is the procedure δ that rejects H_0 if $U \ge T^{-1}_{m+n-2}(1-\alpha_0)$. Theorem 9.6.2 states some useful properties of two-sample t tests analogous to those of Theorem 9.5.1. The proof is so similar to that of Theorem 9.5.1 that we shall not present it here.

Theorem 9.6.2 Level and Unbiasedness of Two-Sample t Tests. Let δ be the two-sample t test defined above. The power function $\pi(\mu_1, \mu_2, \sigma^2|\delta)$ has the following properties:

i. $\pi(\mu_1, \mu_2, \sigma^2|\delta) = \alpha_0$ when $\mu_1 = \mu_2$,
ii. $\pi(\mu_1, \mu_2, \sigma^2|\delta) < \alpha_0$ when $\mu_1 < \mu_2$,
iii. $\pi(\mu_1, \mu_2, \sigma^2|\delta) > \alpha_0$ when $\mu_1 > \mu_2$,
iv. $\pi(\mu_1, \mu_2, \sigma^2|\delta) \to 0$ as $\mu_1 - \mu_2 \to -\infty$,
v. $\pi(\mu_1, \mu_2, \sigma^2|\delta) \to 1$ as $\mu_1 - \mu_2 \to \infty$.

Furthermore, the test δ has size α_0 and is unbiased. ■

Note: The Other One-Sided Hypotheses. If the hypotheses are

$$H_0\colon \mu_1 \ge \mu_2 \quad \text{versus} \quad H_1\colon \mu_1 < \mu_2, \tag{9.6.7}$$

the corresponding level α_0 t test is to reject H_0 when $U \le -T^{-1}_{m+n-2}(1-\alpha_0)$. This test has properties analogous to those of the other one-sided test.

P-values are computed in much the same way as they were for the one-sample t test. The proof of Theorem 9.6.3 is virtually the same as the proof of Theorem 9.5.2 and is not given here.

Theorem 9.6.3 p-Values for Two-Sample t Tests. Suppose that we are testing either the hypotheses in Eq. (9.6.1) or the hypotheses in Eq. (9.6.7). Let u be the observed value of the statistic U in Eq. (9.6.3), and let $T_{m+n-2}(\cdot)$ be the c.d.f. of the t distribution with $m+n-2$ degrees of freedom. Then the p-value for the hypotheses in Eq. (9.6.1) is $1 - T_{m+n-2}(u)$ and the p-value for the hypotheses in Eq. (9.6.7) is $T_{m+n-2}(u)$. ■

Example 9.6.2 Rain from Seeded Clouds. In Example 9.6.1, we actually have 26 observations of unseeded clouds to go with the 26 observations of seeded clouds. Let $X_1, \dots, X_{26}$ be the log-rainfall measurements from the seeded clouds, and let $Y_1, \dots, Y_{26}$ be the measurements from the unseeded clouds. We model all of the measurements as independent with the X_i's having the normal distribution with mean μ_1 and variance σ^2, and the Y_i's having the normal distribution with mean μ_2 and variance σ^2. For now, we model the two distributions as having a common variance. Suppose that we wish to test whether or not the mean log-rainfall from seeded clouds is larger than the mean log-rainfall from unseeded clouds. We choose the null and alternative hypotheses so that type I error corresponds to claiming that seeding increases rainfall when, in fact, it does not increase rainfall. That is, the null hypothesis is $H_0 : \mu_1 \le \mu_2$ and the alternative hypothesis is $H_1 : \mu_1 > \mu_2$. We choose a level of significance of $\alpha_0 = 0.01$. Before proceeding with the formal test, it is a good idea to look at the data first. Figure 9.15 contains histograms of the log-rainfalls of both seeded and unseeded clouds. The two samples look different, with the seeded clouds appearing to have larger log-rainfalls. The formal test requires us to compute the statistics

$$\overline{X}_m = 5.13, \qquad \overline{Y}_n = 3.99,$$
$$S^2_X = 63.96, \quad \text{and} \quad S^2_Y = 67.39.$$

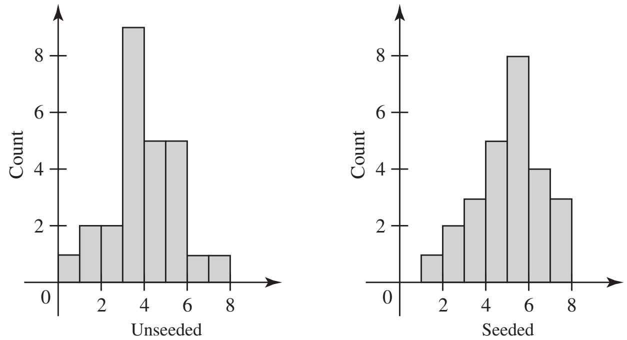


Figure 9.15 Histograms of seeded and unseeded clouds in Example 9.6.2.

The critical value is $T_{50}^{-1}(0.99) = 2.403$, and the test statistic is

$$U = \frac{50^{1/2}(5.13 - 3.99)}{\left(\frac{1}{26} + \frac{1}{26}\right)^{1/2}(63.96 + 67.39)^{1/2}} = 2.544,$$

which is greater than 2.403. So, we would reject the null hypothesis at level of significance $\alpha_0 = 0.01$. The p-value is the smallest level at which we would reject H_0, namely, $1 - T_{50}(2.544) = 0.007$. ◀

Example 9.6.3 Roman Pottery in Britain. Tubb, Parker, and Nickless (1980) describe a study of samples of pottery from the Roman era found in various locations in Great Britain. One measurement made on each sample of pottery was the percentage of the sample that was aluminum oxide. Suppose that we are interested in comparing the aluminum oxide percentages at two different locations. There were $m = 14$ samples analyzed from Llanederyn, with sample average of $\overline{X}_m = 12.56$ and $S_X^2 = 24.65$. Another $n = 5$ samples came from Ashley Rails, with $\overline{Y}_n = 17.32$ and $S_Y^2 = 11.01$. One of the sample sizes is too small for the histogram to be very illuminating. Suppose that we model the data as normal random variables with two different means μ_1 and μ_2 but common variance σ^2. We want to test the null hypothesis $H_0 : \mu_1 \geq \mu_2$ against the alternative hypothesis $H_1 : \mu_1 < \mu_2$. The observed value of U defined by Eq. (9.6.3) is -6.302. From the table of the t distribution in this book, with $m + n - 2 = 17$ degrees of freedom, we find that $T_{17}^{-1}(0.995) = 2.898$ and $U < -2.898$. So, we would reject H_0 at any level $\alpha_0 \geq 0.005$. Indeed, the p-value associated with this value of U is $T_{17}(-6.302) = 4 \times 10^{-6}$. ◀

Power of the Test

For each parameter vector $\theta = (\mu_1, \mu_2, \sigma^2)$, the power function of the two-sample t test can be computed using the noncentral t distribution introduced in Definition 9.5.1. Almost identical reasoning to that which led to Theorem 9.5.3 proves the following.

Theorem 9.6.4 Power of Two-Sample t Test. Assume the conditions stated earlier in this section. Let U be defined in Eq. (9.6.6). Then U has the noncentral t distribution with $m + n - 2$ degrees of freedom and noncentrality parameter

$$\psi = \frac{\mu_1 - \mu_2}{\sigma\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}}. \tag{9.6.8}$$

We can use Fig. 9.12 on page 580 to approximate power calculations if we do not have an appropriate computer program handy.

Example 9.6.4 Roman Pottery in Britain. In Example 9.6.3, if the Llanederyn mean is less than the Ashley Rails mean by 1.5σ, then $\psi = 1.5/(1/14 + 1/5)^{1/2} = 2.88$. The power of a level 0.01 test of $H_0 : \mu_1 \ge \mu_2$ appears to be about 0.65 in the right panel of Fig. 9.12. (The actual power is 0.63.) ◀

Two-Sided Alternatives

The two-sample t test can easily be adapted to testing the following hypotheses at a specified level of significance α_0:

$$H_0: \mu_1 = \mu_2, \quad \text{versus} \quad H_1: \mu_1 \ne \mu_2. \tag{9.6.9}$$

The size α_0 two-sided t test rejects H_0 if $|U| \ge c$ where $c = T^{-1}_{m+n-2}(1 - \alpha_0/2)$, and the statistic U is defined in Eq. (9.6.3). The p-value when $U = u$ is observed equals $2[1 - T_{m+n-2}(|u|)]$. (See Exercise 9.)

Example 9.6.5 Comparing Copper Ores. Suppose that a random sample of eight specimens of ore is collected from a certain location in a copper mine, and the amount of copper in each of the specimens is measured in grams. We shall denote these eight amounts by $X_1, \dots, X_8$ and shall suppose that the observed values are such that $\overline{X}_8 = 2.6$ and $S^2_X = 0.32$. Suppose also that a second random sample of 10 specimens of ore is collected from another part of the mine. We shall denote the amounts of copper in these specimens by $Y_1, \dots, Y_{10}$ and shall suppose that the observed values in grams are such that $\overline{Y}_{10} = 2.3$, and $S^2_Y = 0.22$. Let μ_1 denote the mean amount of copper in all the ore at the first location in the mine, let μ_2 denote the mean amount of copper in all the ore at the second location, and suppose that the hypotheses (9.6.9) are to be tested.

We shall assume that all the observations have a normal distribution, and the variance is the same at both locations in the mine, even though the means may be different. In this example, the sample sizes are $m = 8$ and $n = 10$, and the value of the statistic U defined by Eq. (9.6.3) is 3.442. Also, by the use of a table of the t distribution with 16 degrees of freedom, it is found that $T^{-1}_{16}(0.995) = 2.921$, so that the tail area corresponding to this observed value of U is less than 2×0.005. Hence, the null hypothesis will be rejected for any specified level of significance $\alpha_0 \ge 0.01$. (In fact, the two-sided tail area associated with $U = 3.442$ is 0.003.) ◀

The power function of the two-sided two-sample t test is based on the noncentral t distribution in the same way as was the power function of the one-sample two-sided t test. The test δ that rejects $H_0 : \mu_1 = \mu_2$ when $|U| \ge c$ has power function

$$\pi(\mu_1, \mu_2, \sigma^2|\delta) = T_{m+n-2}(-c|\psi) + 1 - T_{m+n-2}(c|\psi),$$

where $T_{m+n-2}(\cdot|\psi)$ is the c.d.f. of the noncentral t distribution with $m + n - 2$ degrees of freedom and noncentrality parameter ψ given in Eq. (9.6.8). Figure 9.14 on page 583 can be used to approximate the power function if appropriate software is not available.

The Two-Sample t Test as a Likelihood Ratio Test

In this section, we shall show that the two-sample t test for the hypotheses (9.6.1) is a likelihood ratio test. After the values $x_1, \dots, x_m$ and $y_1, \dots, y_n$ in the two samples have been observed, the likelihood function $g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma^2)$ is

$$g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma^2) = f_m(\boldsymbol{x}|\mu_1, \sigma^2) f_n(\boldsymbol{y}|\mu_2, \sigma^2).$$

Here, both $f_m(\boldsymbol{x}|\mu_1, \sigma^2)$ and $f_n(\boldsymbol{y}|\mu_2, \sigma^2)$ have the form given in Eq. (9.5.9), and the value of σ^2 is the same in both terms. In this case, $\Omega_0 = \{(\mu_1, \mu_2, \sigma^2) : \mu_1 \le \mu_2\}$. The likelihood ratio statistic is

$$\Lambda(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sup_{\{(\mu_1,\mu_2,\sigma^2):\mu_1\le\mu_2\}} g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma^2)}{\sup_{(\mu_1,\mu_2,\sigma^2)} g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma^2)}. \tag{9.6.10}$$

The likelihood ratio test procedure then specifies that H_0 should be rejected if $\Lambda(\boldsymbol{x}, \boldsymbol{y}) \le k$, where k is typically chosen so that the test has a desired level α_0.

To facilitate the maximizations in (9.6.10), let

$$s_x^2 = \sum_{i=1}^{m}(x_i - \overline{x}_m)^2, \text{ and } s_y^2 = \sum_{i=1}^{n}(y_i - \overline{y}_n)^2.$$

Then we can write

$$g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{(m+n)/2}} \exp\left(-\frac{1}{2\sigma^2}\left[m(\overline{x}_m - \mu_1)^2 + n(\overline{y}_n - \mu_2)^2 + s_x^2 + s_y^2\right]\right).$$

The denominator of (9.6.10) is maximized by the overall M.L.E.'s, that is, when

$$\mu_1 = \overline{x}_m, \qquad \mu_2 = \overline{y}_n, \qquad \text{and} \quad \sigma^2 = \frac{1}{m+n}(s_x^2 + s_y^2). \tag{9.6.11}$$

For the numerator of (9.6.10), when $\overline{x}_m \le \overline{y}_n$, the parameter vector in (9.6.11) is in Ω_0, and hence the maximum also occurs at the values in Eq. (9.6.11). Hence, $\Lambda(\boldsymbol{x}, \boldsymbol{y}) = 1$ if $\overline{x}_m \le \overline{y}_n$.

For the other case, when $\overline{x}_m > \overline{y}_n$, it is not difficult to see that $\mu_1 = \mu_2$ is required in order to achieve the maximum. In these cases, the maximum occurs when

$$\mu_1 = \mu_2 = \frac{m\overline{x}_m + n\overline{y}_n}{m+n},$$

$$\sigma^2 = \frac{mn(\overline{x}_m - \overline{y}_n)^2/(m+n) + s_x^2 + s_y^2}{m+n}.$$

Substituting all of these values into (9.6.10) yields

$$\Lambda(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} 1 & \text{if } \overline{x}_m \le \overline{y}_n, \\ (1+v^2)^{-(m+n)/2} & \text{if } \overline{x}_m > \overline{y}_n, \end{cases}$$

where

$$v = \frac{(\overline{x}_m - \overline{y}_n)}{\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}(s_x^2 + s_y^2)^{1/2}}. \tag{9.6.12}$$

If $k < 1$, it is straightforward to show that $\Lambda(\boldsymbol{x}, \boldsymbol{y}) \le k$ is equivalent to $v \ge k'$ for some other constant k'. Finally, note that $(m+n-2)^{1/2}v$ is the observed value of U, so

the likelihood ratio test is to reject H_0 when $U \geq c$, for some constant c. This is the same as the two-sample t test. The preceding argument can easily be adapted to handle the other one-sided hypotheses and the two-sided case. (See Exercise 13 for the two-sided case.)

Unequal Variances

Known Ratio of Variances The t test can be extended to a problem in which the variances of the two normal distributions are not equal but the ratio of one variance to the other is known. Specifically, suppose that $X_1, \ldots, X_m$ form a random sample from the normal distribution with mean μ_1 and variance σ_1^2, and $Y_1, \ldots, Y_n$ form an independent random sample from another normal distribution with mean μ_2 and variance σ_2^2. Suppose also that the values of μ_1, μ_2, σ_1^2, and σ_2^2 are unknown but that $\sigma_2^2 = k\sigma_1^2$, where k is a known positive constant. Then it can be shown (see Exercise 4 at the end of this section) that when $\mu_1 = \mu_2$, the following random variable U will have the t distribution with $m + n - 2$ degrees of freedom:

$$U = \frac{(m+n-2)^{1/2}(\overline{X}_m - \overline{Y}_n)}{\left(\frac{1}{m} + \frac{k}{n}\right)^{1/2}\left(S_X^2 + \frac{S_Y^2}{k}\right)^{1/2}}. \tag{9.6.13}$$

Hence, the statistic U defined by Eq. (9.6.13) can be used for testing either the hypotheses (9.6.1) or the hypotheses (9.6.9).

The Behrens-Fisher Problem If the values of all four parameters μ_1, μ_2, σ_1^2, and σ_2^2 are unknown, and if the value of the ratio σ_1^2/σ_2^2 is also unknown, then the problem of testing the hypotheses (9.6.1) or the hypotheses (9.6.9) becomes very difficult. Even the likelihood ratio statistic Λ has no known distribution. This problem is known as the *Behrens-Fisher problem*. Some simulation methods for the Behrens-Fisher problem will be described in Chapter 12 (Examples 12.2.4 and 12.6.10). Various other test procedures have been proposed, but most of them have been the subject of controversy in regard to their appropriateness or usefulness. The most popular of the proposed methods was developed in a series of articles by Welch (1938, 1947, 1951). Welch proposed using the statistic

$$V = \frac{\overline{X}_m - \overline{Y}_n}{\left(\frac{S_X^2}{m(m-1)} + \frac{S_Y^2}{n(n-1)}\right)^{1/2}}. \tag{9.6.14}$$

Even when $\mu_1 = \mu_2$, the distribution of V is not known in closed form. However, Welch approximated the distribution of V by a t distribution as follows. Let

$$W = \frac{S_X^2}{m(m-1)} + \frac{S_Y^2}{n(n-1)}, \tag{9.6.15}$$

and approximate the distribution of W by a gamma distribution with the same mean and variance as W. (See Exercise 12.) If we were now to assume that W actually had this approximating gamma distribution, then V would have the t distribution with

degrees of freedom

$$\frac{\left(\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{\sigma_1^2}{m}\right)^2+\frac{1}{n-1}\left(\frac{\sigma_2^2}{n}\right)^2}. \tag{9.6.16}$$

Next, substitute the unbiased estimates $s_x^2/(m-1)$ and $s_y^2/(n-1)$ for σ_1^2 and σ_2^2, respectively, in (9.6.16) to obtain the degrees of freedom for Welch's t distribution approximation:

$$\nu=\frac{\left(\frac{s_x^2}{m(m-1)}+\frac{s_y^2}{n(n-1)}\right)^2}{\frac{1}{(m-1)^3}\left(\frac{s_x^2}{m}\right)^2+\frac{1}{(n-1)^3}\left(\frac{s_y^2}{n}\right)^2}. \tag{9.6.17}$$

In Eq. (9.6.17), s_x^2 and s_y^2 are the observed values of S_X^2 and S_Y^2. To summarize Welch's procedure, act as if V in Eq. (9.6.14) had the t distribution with ν degrees of freedom when $\mu_1=\mu_2$. Tests of one-sided and two-sided hypotheses are then constructed by comparing V to various quantiles of the t distribution with ν degrees of freedom. If ν is not an integer, round it to the nearest integer or use a computer program that can handle t distributions with noninteger degrees of freedom.

Example 9.6.6

Comparing Copper Ores. Using the data from Example 9.6.5, we compute

$$V=\frac{2.6-2.3}{\left(\frac{0.32}{8\times 7}+\frac{0.22}{10\times 9}\right)^{1/2}}=3.321,$$

$$\nu=\frac{\left(\frac{0.32}{8\times 7}+\frac{0.22}{10\times 9}\right)^2}{\frac{1}{7^3}\left(\frac{0.32}{8}\right)^2+\frac{1}{9^3}\left(\frac{0.22}{10}\right)^2}=12.49.$$

The p-value associated with the observed data for the hypotheses (9.6.9) is $2[1-T_{12.49}(3.321)]=0.0058$, not much different than what we obtained in Example 9.6.5. ◀

Likelihood Ratio Test An alternative to the Welch approximation described above would be to apply the large-sample approximation of Theorem 9.1.4. Using the same notation as earlier in the section, we can write the likelihood function as

$$g(\boldsymbol{x},\boldsymbol{y}|\mu_1,\mu_2,\sigma_1^2,\sigma_2^2) \tag{9.6.18}$$
$$=\frac{1}{(2\pi\sigma_1^2)^{m/2}(2\pi\sigma_2^2)^{n/2}}\exp\left(-\frac{m(\overline{x}_m-\mu_1)^2+s_x^2}{2\sigma_1^2}-\frac{n(\overline{y}_n-\mu_2)^2+s_y^2}{2\sigma_2^2}\right).$$

The overall M.L.E.'s are

$$\hat{\mu}_1=\overline{x}_m,\ \hat{\mu}_2=\overline{y}_n,\ \hat{\sigma}_1^2=\frac{s_x^2}{m},\ \hat{\sigma}_2^2=\frac{s_y^2}{n}. \tag{9.6.19}$$

Under $H_0: \mu_1 = \mu_2$, we cannot find formulas for the M.L.E.'s. However, if we let $\hat{\mu}$ stand for the common value of $\hat{\mu}_1 = \hat{\mu}_2$, we find that the M.L.E.'s satisfy the following equations:

$$\hat{\sigma}_1^2 = \frac{1}{m}\left[s_x^2 + m(\overline{x}_m - \hat{\mu})^2\right], \tag{9.6.20}$$

$$\hat{\sigma}_2^2 = \frac{1}{n}\left[s_y^2 + n(\overline{y}_n - \hat{\mu})^2\right], \tag{9.6.21}$$

$$\hat{\mu} = \frac{\dfrac{m\overline{x}_m}{\hat{\sigma}_1^2} + \dfrac{n\overline{y}_n}{\hat{\sigma}_2^2}}{\dfrac{m}{\hat{\sigma}_1^2} + \dfrac{n}{\hat{\sigma}_2^2}}. \tag{9.6.22}$$

These equations can be solved recursively even though we do not have a closed-form solution. One algorithm is the following:

1. Set $k = 0$ and pick a starting value $\hat{\mu}^{(0)}$, such as $(m\overline{x}_m + n\overline{y}_n)/(m + n)$.
2. Compute $\hat{\sigma}_1^{2(k)}$ and $\hat{\sigma}_2^{2(k)}$ by substituting $\hat{\mu}^{(k)}$ into Eqs. (9.6.20) and (9.6.21).
3. Compute $\hat{\mu}^{(k+1)}$ by substituting $\hat{\sigma}_1^{2(k)}$ and $\hat{\sigma}_2^{2(k)}$ into Eq. (9.6.22).
4. If $\hat{\mu}^{(k+1)}$ is close enough to $\hat{\mu}^{(k)}$ stop. Otherwise, replace k by $k + 1$ and return to step 2.

Example 9.6.7

Comparing Copper Ores. Using the data in Example 9.6.5, we will start with $\hat{\mu}^{(0)} = (8 \times 2.6 + 10 \times 2.3)/18 = 2.433$. Plugging this value into Eqs. (9.6.20) and (9.6.21) gives us $\hat{\sigma}_1^{2(0)} = 0.068$ and $\hat{\sigma}_2^{2(0)} = 0.0398$. Plugging these into Eq. (9.6.22) gives $\hat{\mu}^{(1)} = 2.396$. After 13 iterations the values stop changing and our final M.L.E.'s are $\hat{\mu} = 2.347$, $\hat{\sigma}_1^2 = 0.1039$, and $\hat{\sigma}_2^2 = 0.0242$. We can then substitute these M.L.E.'s into the likelihood function (9.6.18) to get the numerator of the likelihood ratio statistic $\Lambda(\boldsymbol{x}, \boldsymbol{y})$. (Remember to substitute $\hat{\mu}$ for both μ_1 and μ_2.) We can also substitute the overall M.L.E.'s (9.6.19) into (9.6.18) to get the denominator of $\Lambda(\boldsymbol{x}, \boldsymbol{y})$. The result is $\Lambda(\boldsymbol{x}, \boldsymbol{y}) = 0.01356$. Theorem 9.1.4 says that we should compare $-2 \log \Lambda(\boldsymbol{x}, \boldsymbol{y}) = 8.602$ to a critical value of the χ^2 distribution with one degree of freedom. The p-value associated with the observed statistic is the probability that a χ^2 random variable with one degree of freedom is greater than 8.602, namely, 0.003. This is the same as the p-value that we obtained in Example 9.6.5 when we assumed that the two variances were the same. ◀

For the cases of one-sided hypotheses such as (9.6.1) and (9.6.7), the likelihood ratio statistic is a bit more complicated. For example, if $\mu_1 = \mu_2$, $-2 \log \Lambda(\boldsymbol{X}, \boldsymbol{Y})$ converges in distribution to a distribution that is neither discrete nor continuous. We will not discuss this case further in this book.

Summary

Suppose that we observe independent random samples from two normal distributions: $X_1, \ldots, X_m$ having mean μ_1 and variance σ_1^2, and $Y_1, \ldots, Y_n$ having mean μ_2 and variance σ_2^2. For testing hypotheses about μ_1 and μ_2, t tests are available if we assume that $\sigma_1^2 = \sigma_2^2$. The t tests all make use of the statistic U defined in Eq. (9.6.3). To test $H_0: \mu_1 = \mu_2$ versus $H_1: \mu_1 \neq \mu_2$ at level α_0, reject H_0 if $|U| \geq$

$T^{-1}_{m+n-2}(1-\alpha_0/2)$, where T^{-1}_{m+n-2} is the quantile function of the t distribution with $m+n-2$ degrees of freedom. To test $H_0: \mu_1 \le \mu_2$ versus $H_1: \mu_1 > \mu_2$ at level α_0, reject H_0 if $U > T^{-1}_{m+n-2}(1-\alpha_0)$. To test $H_0: \mu_1 \ge \mu_2$ versus $H_1: \mu_1 < \mu_2$ at level α_0, reject H_0 if $U < -T^{-1}_{m+n-2}(1-\alpha_0)$. The power functions of these tests can be computed using the family of noncentral t distributions. Approximate tests are available if we do not assume that $\sigma_1^2 = \sigma_2^2$.

Exercises

1. In Example 9.6.3, we discussed Roman pottery found at two different locations in Great Britain. There were samples found at other locations as well. One other location, Island Thorns, had five samples $X_1, \ldots, X_n$ with an average aluminum oxide percentage of $\overline{X} = 18.18$ with $\sum_{i=1}^{5}(X_i - \overline{X})^2 = 12.61$. Let $Y_1, \ldots, Y_5$ be the five sample measurements from Ashley Rails in Example 9.6.3. Test the null hypothesis that the mean aluminum oxide percentages at Ashely Rails and Island Thorns are the same versus the alternative that they are different at level $\alpha_0 = 0.05$.

2. Suppose that a certain drug A was administered to eight patients selected at random, and after a fixed time period, the concentration of the drug in certain body cells of each patient was measured in appropriate units. Suppose that these concentrations for the eight patients were found to be as follows:

1.23, 1.42, 1.41, 1.62, 1.55, 1.51, 1.60, and 1.76.

Suppose also that a second drug B was administered to six different patients selected at random, and when the concentration of drug B was measured in a similar way for these six patients, the results were as follows:

1.76, 1.41, 1.87, 1.49, 1.67, and 1.81.

Assuming that all the observations have a normal distribution with a common unknown variance, test the following hypotheses at the level of significance 0.10: The null hypothesis is that the mean concentration of drug A among all patients is at least as large as the mean concentration of drug B. The alternative hypothesis is that the mean concentration of drug B is larger than that of drug A.

3. Consider again the conditions of Exercise 2, but suppose now that it is desired to test the following hypotheses: The null hypothesis is that the mean concentration of drug A among all patients is the same as the mean concentration of drug B. The alternative hypothesis, which is two-sided, is that the mean concentrations of the two drugs are not the same. Find the number c so that the level 0.05 two-sided t test will reject H_0 when $|U| \ge c$, where U is defined by Eq. (9.6.3). Also, perform the test.

4. Suppose that $X_1, \ldots, X_m$ form a random sample from the normal distribution with mean μ_1 and variance σ_1^2, and $Y_1, \ldots, Y_n$ form an independent random sample from the normal distribution with mean μ_2 and variance σ_2^2. Show that if $\mu_1 = \mu_2$ and $\sigma_2^2 = k\sigma_1^2$, then the random variable U defined by Eq. (9.6.13) has the t distribution with $m+n-2$ degrees of freedom.

5. Consider again the conditions and observed values of Exercise 2. However, suppose now that each observation for drug A has an unknown variance σ_1^2, and each observation for drug B has an unknown variance σ_2^2, but it is known that $\sigma_2^2 = (6/5)\sigma_1^2$. Test the hypotheses described in Exercise 2 at the level of significance 0.10.

6. Suppose that $X_1, \ldots, X_m$ form a random sample from the normal distribution with unknown mean μ_1 and unknown variance σ^2, and $Y_1, \ldots, Y_n$ form an independent random sample from another normal distribution with unknown mean μ_2 and the same unknown variance σ^2. For each constant λ $(-\infty < \lambda < \infty)$, construct a t test of the following hypotheses with $m+n-2$ degrees of freedom:

$$
\begin{aligned}
H_0:&\quad \mu_1 - \mu_2 = \lambda,\\
H_1:&\quad \mu_1 - \mu_2 \ne \lambda.
\end{aligned}
$$

7. Consider again the conditions of Exercise 2. Let μ_1 denote the mean of each observation for drug A, and let μ_2 denote the mean of each observation for drug B. It is assumed, as in Exercise 2, that all the observations have a common unknown variance. Use the results of Exercise 6 to construct a confidence interval for $\mu_1 - \mu_2$ with confidence coefficient 0.90.

8. In Example 9.6.5, determine the power of a level 0.01 test if $|\mu_1 - \mu_2| = \sigma$.

9. Suppose that we wish to test the hypotheses (9.6.9). We shall use the statistic U defined in Eq. (9.6.3) and reject H_0 if $|U|$ is large. Prove that the p-value when $U = u$ is observed is $2[1 - T_{m+n-2}(|u|)]$.

10. Lyle et al. (1987) ran an experiment to study the effect of a calcium supplement on the blood pressure of African American males. A group of 10 men received a calcium supplement, and another group of 11 men received a placebo. The experiment lasted 12 weeks. Both

Table 9.2 Blood pressure data for Exercise 10

Calcium	7	−4	18	17	−3	−5	1	10	11	−2	
Placebo	−1	12	−1	−3	3	−5	5	2	−11	−1	−3

before and after the 12-week period, each man had his systolic blood pressure measured while at rest. The changes (after minus before) are given in Table 9.2. Test the null hypothesis that the mean change in blood pressure for the calcium supplement group is lower than the mean change in blood pressure for the placebo group. Use level $\alpha_0 = 0.1$.

11. Frisby and Clatworthy (1975) studied the times that it takes subjects to fuse random-dot stereograms. Random-dot stereograms are pairs of images that appear at first to be random dots. After a subject looks at the pair of images from the proper distance and her eyes cross just the right amount, a recognizable object appears from the fusion of the two images. The experimenters were concerned with the extent to which prior information about the recognizable object affected the time it took to fuse the images.

One group of 43 subjects was not shown a picture of the object before being asked to fuse the images. Their average time was $\overline{X}_{43} = 8.560$ and $S_X^2 = 2745.7$. The second group of 35 subjects was shown a picture of the object, and their sample statistics were $\overline{Y}_{35} = 5.551$ and $S_Y^2 = 783.9$. The null hypothesis is that the mean time of the first group is no larger than the mean time of the second group, while the alternative hypothesis is that the first group takes longer.

a. Test the hypotheses at the level of significance $\alpha_0 = 0.01$, assuming that the variances are equal for the two groups.

b. Test the hypotheses at the level of significance $\alpha_0 = 0.01$, using Welch's approximate test.

12. Find the mean a and variance b of the random variable W in Eq. (9.6.15). Now, let a and b be the mean and variance, respectively, of the gamma distribution with parameters α and β. Prove that 2α equals the expression in (9.6.16).

13. Let U be as defined in Eq. (9.6.3), and suppose that it is desired to test the hypotheses in Eq. (9.6.9). Prove that each likelihood ratio test has the following form: reject H_0 if $|U| \ge c$, where c is a constant. *Hint:* First prove that $\Lambda(\boldsymbol{x}, \boldsymbol{y}) = (1 + v^2)^{-(m+n)/2}$, where v was defined in Eq. (9.6.12).

9.7 The F Distributions

In this section, we introduce the family of F distributions. This family is useful in two different hypothesis-testing situations. The first situation is when we wish to test hypotheses about the variances of two different normal distributions. These tests, which we shall derive in this section, are based on a statistic that has an F distribution. The second situation will arise in Chapter 11 when we test hypotheses concerning the means of more than two normal distributions.

Definition of the F Distribution

Example 9.7.1 Rain from Seeded Clouds. In Example 9.6.1, we were interested in comparing the distributions of log-rainfalls from seeded and unseeded clouds. In Example 9.6.2, we used the two-sample t test to compare the means of these distributions under the assumption that the variances of the two distributions were the same. It would be good to have a procedure for testing whether or not such an assumption is warranted. ◀

In this section, we shall introduce a family of distributions, called the F distributions, that arises in many important problems of testing hypotheses in which two or more normal distributions are to be compared on the basis of random samples from

each of the distributions. In particular, it arises naturally when we wish to compare the variances of two normal distributions.

Definition 9.7.1 The F distributions. Let Y and W be independent random variables such that Y has the χ^2 distribution with m degrees of freedom and W has the χ^2 distribution with n degrees of freedom, where m and n are given positive integers. Define a new random variable X as follows:

$$X = \frac{Y/m}{W/n} = \frac{nY}{mW}. \tag{9.7.1}$$

Then the distribution of X is called the *F distribution with m and n degrees of freedom*.

Theorem 9.7.1 gives the general p.d.f. of an F distribution. Its proof relies on the methods of Sec. 3.9 and will be postponed until the end of this section.

Theorem 9.7.1 Probability Density Function. Let X have the F distribution with m and n degrees of freedom. Then its p.d.f. $f(x)$ is as follows, for $x > 0$:

$$f(x) = \frac{\Gamma\left[\frac{1}{2}(m+n)\right] m^{m/2} n^{n/2}}{\Gamma\left(\frac{1}{2}m\right)\Gamma\left(\frac{1}{2}n\right)} \cdot \frac{x^{(m/2)-1}}{(mx+n)^{(m+n)/2}}, \tag{9.7.2}$$

and $f(x) = 0$ for $x \le 0$.

Properties of the F Distributions

When we speak of the F distribution with m and n degrees of freedom, the order in which the numbers m and n are given is important, as can be seen from the definition of X in Eq. (9.7.1). When $m \neq n$, the F distribution with m and n degrees of freedom and the F distribution with n and m degrees of freedom are two different distributions. Theorem 9.7.2 gives a result relating the two distributions just mentioned along with a relationship between F distributions and t distributions.

Theorem 9.7.2 If X has the F distribution with m and n degrees of freedom, then its reciprocal $1/X$ has the F distribution with n and m degrees of freedom. If Y has the t distribution with n degrees of freedom, then Y^2 has the F distribution with 1 and n degrees of freedom.

Proof The first statement follows from the representation of X as the ratio of two random variables, in Definition 9.7.1. The second statement follows from the representation of a t random variable in the form of Eq. (8.4.1). ■

Two short tables of quantiles for F distributions are given at the end of this book. In these tables, we give only the 0.95 quantile and the 0.975 quantile for different possible pairs of values of m and n. In other words, if G denotes the c.d.f. of the F distribution with m and n degrees of freedom, then the tables give the values of x_1 and x_2 such that $G(x_1) = 0.95$ and $G(x_2) = 0.975$. By applying Theorem 9.7.2, it is possible to use the tables to obtain the 0.05 and 0.025 quantiles of an F distribution. Most statistical software will compute the c.d.f. and quatiles for general F distributions.

Example 9.7.2 Determining the 0.05 Quantile of an F Distribution. Suppose that a random variable X has the F distribution with 6 and 12 degrees of freedom. We shall determine the 0.05 quantile of X, that is, the value of x such that $\Pr(X < x) = 0.05$.

If we let $Y = 1/X$, then Y will have the F distribution with 12 and 6 degrees of freedom. It can be found from the table given at the end of this book that $\Pr(Y \le 4.00) = 0.95$; hence, $\Pr(Y > 4.00) = 0.05$. Since $Y > 4.00$ if and only if $X < 0.25$, it follows that $\Pr(X < 0.25) = 0.05$. Because F distributions are continuous, $\Pr(X \le 0.25) = 0.05$, and 0.25 is the 0.05 quantile of X. ◀

Comparing the Variances of Two Normal Distributions

Suppose that the random variables $X_1, \dots, X_m$ form a random sample of m observations from a normal distribution for which both the mean μ_1 and the variance σ_1^2 are unknown, and suppose also that the random variables $Y_1, \dots, Y_n$ form an independent random sample of n observations from another normal distribution for which both the mean μ_2 and the variance σ_2^2 are unknown. Suppose finally that the following hypotheses are to be tested at a specified level of significance α_0 $(0 < \alpha_0 < 1)$:

$$\begin{aligned} H_0&: \quad \sigma_1^2 \le \sigma_2^2, \\ H_1&: \quad \sigma_1^2 > \sigma_2^2. \end{aligned} \tag{9.7.3}$$

For each test procedure δ, we shall let $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta)$ denote the power function of δ. Later in this section, we shall derive the likelihood ratio test. For now, define S_X^2 and S_Y^2 to be the sums of squares defined in Eq. (9.6.2). Then $S_X^2/(m-1)$ and $S_Y^2/(n-1)$ are estimators of σ_1^2 and σ_2^2, respectively. It makes intuitive sense that we should reject H_0 if the ratio of these two estimators is large. That is, define

$$V = \frac{S_X^2/(m-1)}{S_Y^2/(n-1)}, \tag{9.7.4}$$

and reject H_0 if $V \ge c$, where c is chosen to make the test have a desired level of significance.

Definition 9.7.2 *F* test. The test procedure defined above is called an F *test*.

Properties of F Tests

Theorem 9.7.3 Distribution of V. Let V be the statistic in Eq. (9.7.4). The distribution of $(\sigma_2^2/\sigma_1^2)V$ is the F distribution with $m-1$ and $n-1$ degrees of freedom. In particular, if $\sigma_1^2 = \sigma_2^2$, then the distribution of V itself is the F distribution with $m-1$ and $n-1$ degrees of freedom.

Proof We know from Theorem 8.3.1 that the random variable S_X^2/σ_1^2 has the χ^2 distribution with $m-1$ degrees of freedom, and the random variable S_Y^2/σ_2^2 has the χ^2 distribution with $n-1$ degrees of freedom. Furthermore, these two random variables are independent, since they are calculated from two independent samples. Therefore, the following random variable V^* has the F distribution with $m-1$ and $n-1$ degrees of freedom:

$$V^* = \frac{S_X^2/[(m-1)\sigma_1^2]}{S_Y^2/[(n-1)\sigma_2^2]}. \tag{9.7.5}$$

It can be seen from Eqs. (9.7.4) and (9.7.5) that $V^* = (\sigma_2^2/\sigma_1^2)V$. This proves the first claim in the theorem. If $\sigma_1^2 = \sigma_2^2$, then $V = V^*$, which proves the second claim. ■

If $\sigma_1^2 = \sigma_2^2$, it is possible to use a table of the F distribution to choose a constant c such that $\Pr(V \ge c) = \alpha_0$, regardless of the common value of σ_1^2 and σ_2^2, and regardless of the values of μ_1 and μ_2. In fact, c will be the $1-\alpha_0$ quantile of the corresponding F distribution. We prove next that the test that rejects H_0 in (9.7.3) if $V \ge c$ has level α_0.

Theorem 9.7.4 Level, Power Function, and P-Values. Let V be the statistic defined in Eq. (9.7.4). Let c be the $1-\alpha_0$ quantile of the F distribution with $m-1$ and $n-1$ degrees of freedom, and let $G_{m-1,n-1}$ be the c.d.f. of that F distribution. Let δ be test that rejects H_0 in (9.7.3) when $V \ge c$. The power function $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta)$ satisfies the following properties:

i. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) = 1 - G_{m-1,n-1}\left(\frac{\sigma_2^2}{\sigma_1^2}c\right)$,

ii. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) = \alpha_0$ when $\sigma_1^2 = \sigma_2^2$,

iii. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) < \alpha_0$ when $\sigma_1^2 < \sigma_2^2$,

iv. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) > \alpha_0$ when $\sigma_1^2 > \sigma_2^2$,

v. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) \to 0$ as $\sigma_1^2/\sigma_2^2 \to 0$,

vi. $\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) \to 1$ as $\sigma_1^2/\sigma_2^2 \to \infty$.

The test δ has level α_0 and is unbiased. The p-value when $V = v$ is observed equals $1 - G_{m-1,n-1}(v)$.

Proof The power function is the probability of rejecting H_0, i.e., the probability that $V \ge c$. Let V^* be as defined in Eq. (9.7.5) so that V^* has the F distribution with $m-1$ and $n-1$ degrees of freedom. Then

$$\begin{aligned}\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) &= \Pr(V \ge c) = \Pr\left(\frac{\sigma_1^2}{\sigma_2^2}V^* \ge c\right) = \Pr\left(V^* \ge \frac{\sigma_2^2}{\sigma_1^2}c\right) \\ &= 1 - G_{m-1,n-1}\left(\frac{\sigma_2^2}{\sigma_1^2}c\right),\end{aligned} \tag{9.7.6}$$

which proves property (i). Property (ii) follows from Theorem 9.7.3. For property (iii), let $\sigma_1^2 < \sigma_2^2$ in Eq. (9.7.6). Since $(\sigma_2^2/\sigma_1^2)c > c$, the expression on the far right of (9.7.6) is less than $1 - G_{m-1,n-1}(c) = \alpha_0$. Similarly, if $\sigma_1^2 > \sigma_2^2$, the expression on the far right of (9.7.6) is greater than $1 - G_{m-1,n-1}(c) = \alpha_0$, proving property (iv). Properties (v) and (vi) follow from property (i) and elementary properties of c.d.f.'s, namely, Property 3.3.2. The fact that δ has level α_0 follows from properties (ii) and (iii). The fact that δ is unbiased follows from properties (ii) and (iv). Finally, the p-value is the smallest α_0 such that we would reject H_0 at level α_0 if $V = v$ were observed. We reject H_0 at level α_0 if and only if $v \ge G^{-1}_{m-1,n-1}(1-\alpha_0)$, which is equivalent to $\alpha_0 \ge 1 - G_{m-1,n-1}(v)$. Hence, $1 - G_{m-1,n-1}(v)$ is the smallest α_0 such that we would reject H_0. ■

Example 9.7.3 Performing an $\boldsymbol{F}$ Test. Suppose that six observations $X_1, \dots, X_6$ are selected at random from a normal distribution for which both the mean μ_1 and the variance σ_1^2 are unknown, and it is found that $S_X^2 = 30$. Suppose also that 21 observations, $Y_1, \dots, Y_{21}$, are selected at random from another normal distribution for which both the mean μ_2 and the variance σ_2^2 are unknown, and that it is found that $S_Y^2 = 40$. We shall carry out an F test of the hypotheses (9.7.3).

In this example, $m = 6$ and $n = 21$. Therefore, when H_0 is true, the statistic V defined by Eq. (9.7.4) will have the F distribution with 5 and 20 degrees of freedom. It follows from Eq. (9.7.4) that the value of V for the given samples is

$$V = \frac{30/5}{40/20} = 3.$$

It is found from the tables given at the end of this book that the 0.95 quantile of the F distribution with 5 and 20 degrees of freedom is 2.71, and the 0.975 quantile of that distribution is 3.29. Hence, the tail area corresponding to the value $V = 3$ is less than 0.05 and greater than 0.025. The hypothesis H_0 that $\sigma_1^2 \le \sigma_2^2$ would therefore be rejected at the level of significance $\alpha_0 = 0.05$, and H_0 would not be rejected at the level of significance $\alpha_0 = 0.025$. (Using a computer program to evaluate the c.d.f. of an F distribution provides the p-value equal to 0.035.) Finally, suppose that it is important to reject H_0 if σ_1^2 is three times as large as σ_2^2. We would then want the power function to be high when $\sigma_1^2 = 3\sigma_2^2$. We use a computer program to compute

$$1 - F_{5,20}\left(2.71 \times \frac{1}{3}\right) = 0.498.$$

Even if σ_1^2 is three times as large as σ_2^2, the level 0.05 test only has about a 50 percent chance of rejecting H_0. ◀

Two-Sided Alternative

Suppose that we wish to test the hypotheses

$$\begin{aligned} H_0&: \quad \sigma_1^2 = \sigma_2^2, \\ H_1&: \quad \sigma_1^2 \ne \sigma_2^2. \end{aligned} \tag{9.7.7}$$

It would make sense to reject H_0 if either $V \le c_1$ or $V \ge c_2$, where V is defined in Eq. (9.7.4) and c_1 and c_2 are constants such that $\Pr(V \le c_1) + \Pr(V \ge c_2) = \alpha_0$ when $\sigma_1^2 = \sigma_2^2$. The most convenient choice of c_1 and c_2 is the one that makes $\Pr(V \le c_1) = \Pr(V \ge c_2) = \alpha_0/2$. That is, choose c_1 and c_2 to be the $\alpha_0/2$ and $1 - \alpha_0/2$ quantiles of the appropriate F distribution.

Example 9.7.4

Rain from Seeded Clouds. In Example 9.6.2, we compared the means of log-rainfalls from seeded and unseeded clouds under the assumption that the two variances were the same. We can now test the null hypothesis that the two variances are the same against the alternative hypothesis that the two variances are different at level of significance $\alpha_0 = 0.05$. Using the statistics given in Example 9.6.2, the value of V is $63.96/67.39 = 0.9491$, since $m = n$. We need to compare this to the 0.025 and 0.975 quantiles of the F distribution with 25 and 25 degrees of freedom. Since our table of F distribution quantiles does not have rows or columns for 25 degrees of freedom, we can either interpolate between 20 and 30 degrees of freedom or use a computer program to compute these quantiles. The quantiles are 0.4484 and 2.2303. Since V is between these two numbers, we would not reject the null hypothesis at level $\alpha_0 = 0.05$. ◀

When $m \ne n$, the two-sided F test constructed above is not unbiased. (See Exercise 19.) Also, if $m \ne n$, it is not possible to write the two-sided F test described above in the form "reject the null hypothesis if $T \ge c$" using the same statistic T for each significance level α_0. Nevertheless, we can still compute the smallest α_0 such

that the two-sided F test with level of significance α_0 would reject H_0. The proof of the following result is left to Exercise 15 in this section.

Theorem 9.7.5 *P*-Value of Equal-Tailed Two-Sided *F* Test. Let V be as defined in (9.7.4). Suppose that we wish to test the hypotheses (9.7.7). Let δ_{α_0} be the equal-tailed two-sided F test that rejects H_0 when $V \le c_1$ or $V \ge c_2$, where c_1 and c_2 are, respectively, the $\alpha_0/2$ and $1-\alpha_0/2$ quantiles of the appropriate F distribution. Then the smallest α_0 such that δ_{α_0} rejects H_0 when $V = v$ is observed is

$$2\min\{1 - G_{m-1,\,n-1}(v),\, G_{m-1,\,n-1}(v)\}. \tag{9.7.8}$$

The F Test as a Likelihood Ratio Test

Next, we shall show that the F test for hypotheses (9.7.3) is a likelihood ratio test. After the values $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ in the two samples have been observed, the likelihood function $g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ is

$$g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = f_m(\boldsymbol{x}|\mu_1, \sigma_1^2) f_n(\boldsymbol{y}|\mu_2, \sigma_2^2).$$

Here, both $f_m(\boldsymbol{x}|\mu_1, \sigma_1^2)$ and $f_n(\boldsymbol{y}|\mu_2, \sigma_2^2)$ have the general form given in Eq. (9.5.9). For the hypotheses in (9.7.3), Ω_0 contains all parameters $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ with $\sigma_1^2 \le \sigma_2^2$, and Ω_1 contains all θ with $\sigma_1^2 > \sigma_2^2$. The likelihood ratio statistic is

$$\Lambda(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sup_{\{(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2):\sigma_1^2\le\sigma_2^2\}} g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)}{\sup_{(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2)} g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)}. \tag{9.7.9}$$

The likelihood ratio test then specifies that H_0 should be rejected if $\Lambda(\boldsymbol{x}, \boldsymbol{y}) \le k$, where k is typically chosen to make the test have a desired level α_0.

To facilitate the maximizations in (9.7.9), let

$$s_x^2 = \sum_{i=1}^{m}(x_i - \overline{x}_m)^2, \text{ and } s_y^2 = \sum_{i=1}^{n}(y_i - \overline{y}_n)^2.$$

Then we can write

$$g(\boldsymbol{x}, \boldsymbol{y}|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \frac{1}{(2\pi)^{(m+n)/2}\sigma_1^m\sigma_2^n}\exp\left(-\frac{1}{2\sigma_1^2}\left[n(\overline{x}_m - \mu_1)^2 + s_x^2\right] - \frac{1}{2\sigma_2^2}\left[n(\overline{y}_n - \mu_2)^2 + s_y^2\right]\right).$$

For both the numerator and denominator of (9.7.9), we need $\mu_1 = \overline{x}_m$ and $\mu_2 = \overline{y}_n$ in order to maximize the likelihood. If $s_x^2/m \le s_y^2/n$, then the numerator is maximized at $\sigma_1^2 = s_x^2/m$ and $\sigma_2^2 = s_y^2/n$. These values also maximize the denominator. Hence, $\Lambda(\boldsymbol{x}, \boldsymbol{y}) = 1$ if $s_x^2/m \le s_y^2/n$. For the other case (the numerator when $s_x^2/m > s_y^2/n$), it is straightforward to show that $\sigma_1^2 = \sigma_2^2$ is required in order to achieve the maximum. In these cases, the maximum occurs when

$$\sigma_1^2 = \sigma_2^2 = \frac{s_x^2 + s_y^2}{m+n}.$$

Substituting all of these values into (9.7.9) yields

$$\Lambda(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} 1 & \text{if } s_x^2/m \le s_y^2/n, \\ d w^{m/2}(1-w)^{n/2} & \text{if } s_x^2/m > s_y^2/m, \end{cases}$$

where

$$w = \frac{s_x^2}{s_x^2 + s_y^2}, \quad \text{and} \quad d = \frac{(m+n)^{(m+n)/2}}{m^{m/2}n^{n/2}}.$$

Note that $s_x^2/m \le s_y^2/n$ if and only if $w \le m/(m+n)$. Next, use the fact that the function $h(w) = w^{m/2}(1-w)^{n/2}$ is decreasing for $m/(m+n) < w < 1$. Finally, note that $h(m/[m+n]) = 1/d$. For $k < 1$, it follows that $\Lambda(\boldsymbol{x}, \boldsymbol{y}) \le k$ if and only if $w \ge k'$ for some other constant k'. This, in turn, is equivalent to $s_x^2/s_y^2 \ge k''$. Since s_x^2/s_y^2 is a positive constant times the observed value of V, the likelihood ratio test rejects H_0 when V is large. This is the same as the F test.

One can easily adapt the above argument for the case in which the inequalities are reversed in the hypotheses. When the hypotheses are (9.7.7), that is, the alternative is two-sided, one can show (see Exercise 16) that the size α_0 likelihood ratio test will reject H_0 if either $V \le c_1$ or $V \ge c_2$. Unfortunately, it is usually tedious to compute the necessary values c_1 and c_2. For this reason, people often abandon the strict likelihood ratio criterion in this case and simply let c_1 and c_2 be the $\alpha_0/2$ and $1-\alpha_0/2$ quantiles of the appropriate F distribution.

Derivation of the p.d.f. of an F distribution

Since the random variables Y and W in Definition 9.7.1 are independent, their joint p.d.f. $g(y, w)$ is the product of their individual p.d.f.'s. Furthermore, since both Y and W have χ^2 distributions, it follows from the p.d.f. of the χ^2 distribution, as given in Eq. 8.2.1, that $g(y, w)$ has the following form, for $y > 0$ and $w > 0$:

$$g(y, w) = cy^{(m/2)-1}w^{(n/2)-1}e^{-(y+w)/2}, \tag{9.7.10}$$

where

$$c = \frac{1}{2^{(m+n)/2}\Gamma\left(\frac{1}{2}m\right)\Gamma\left(\frac{1}{2}n\right)}. \tag{9.7.11}$$

We shall now change variables from Y and W to X and W, where X is defined by Eq. (9.7.1). The joint p.d.f. $h(x, w)$ of X and W is obtained by first replacing y in Eq. (9.7.10) with its expression in terms of x and w and then multiplying the result by $|\partial y/\partial x|$. It follows from Eq. (9.7.1) that $y = (m/n)xw$ and $\partial y/\partial x = (m/n)w$. Hence, the joint p.d.f. $h(x, w)$ has the following form, for $x > 0$ and $w > 0$:

$$h(x, w) = c\left(\frac{m}{n}\right)^{m/2} x^{(m/2)-1}w^{[(m+n)/2]-1}\exp\left[-\frac{1}{2}\left(\frac{m}{n}x + 1\right)w\right]. \tag{9.7.12}$$

Here, the constant c is again given by Eq. (9.7.11).

The marginal p.d.f $f(x)$ of X can be obtained for each value of $x > 0$ from the relation

$$f(x) = \int_0^\infty h(x, w)\,dw. \tag{9.7.13}$$

It follows from Theorem 5.7.3 that

$$\int_0^\infty w^{[(m+n)/2]-1}\exp\left[-\frac{1}{2}\left(\frac{m}{n}x+1\right)w\right]dw = \frac{\Gamma\left[\frac{1}{2}(m+n)\right]}{\left[\frac{1}{2}\left(\frac{m}{n}x+1\right)\right]^{(m+n)/2}}. \tag{9.7.14}$$

From Eqs. (9.7.11) to (9.7.14), we can conclude that the p.d.f $f(x)$ has the form given in Eq. (9.7.2).

Summary

If Y and W are independent with Y having the χ^2 distribution with m degrees of freedom and W having the χ^2 distribution with n degrees of freedom, then $(Y/m)/(W/n)$ has the F distribution with m and n degrees of freedom. Suppose that we observe two independent random samples from two normal distributions with possibly different variances. The ratio V of the usual unbiased estimators of the two variances will have an F distribution when the two variances are equal. Tests of hypotheses about the two variances can be constructed by comparing V to various quantiles of F distributions.

Exercises

1. Consider again the situation described in Exercise 11 of Sec. 9.6. Test the null hypothesis that the variance of the fusion time for subjects who saw a picture of the object is no smaller than the variance for subjects who did see a picture. The alternative hypothesis is that the variance for subjects who saw a picture is smaller than the variance for subjects who did not see a picture. Use a level of significance of 0.05.

2. Suppose that a random variable X has the F distribution with three and eight degrees of freedom. Determine the value of c such that $\Pr(X > c) = 0.975$.

3. Suppose that a random variable X has the F distribution with one and eight degrees of freedom. Use the table of the t distribution to determine the value of c such that $\Pr(X > c) = 0.3$.

4. Suppose that a random variable X has the F distribution with m and n degrees of freedom $(n > 2)$. Show that $E(X) = n/(n-2)$. *Hint:* Find the value of $E(1/Z)$, where Z has the χ^2 distribution with n degrees of freedom.

5. What is the value of the median of the F distribution with m and n degrees of freedom when $m = n$?

6. Suppose that a random variable X has the F distribution with m and n degrees of freedom. Show that the random variable $mX/(mX + n)$ has the beta distribution with parameters $\alpha = m/2$ and $\beta = n/2$.

7. Consider two different normal distributions for which both the means μ_1 and μ_2 and the variances σ_1^2 and σ_2^2 are unknown, and suppose that it is desired to test the following hypotheses:

$$H_0: \quad \sigma_1^2 \le \sigma_2^2,$$
$$H_1: \quad \sigma_1^2 > \sigma_2^2.$$

Suppose further that a random sample consisting of 16 observations for the first normal distribution yields the values $\sum_{i=1}^{16} X_i = 84$ and $\sum_{i=1}^{16} X_i^2 = 563$, and an independent random sample consisting of 10 observations from the second normal distribution yields the values $\sum_{i=1}^{10} Y_i = 18$ and $\sum_{i=1}^{10} Y_i^2 = 72$.

a. What are the M.L.E.'s of σ_1^2 and σ_2^2?

b. If an F test is carried out at the level of significance 0.05, is the hypothesis H_0 rejected or not?

8. Consider again the conditions of Exercise 7, but suppose now that it is desired to test the following hypotheses:

$$H_0: \quad \sigma_1^2 \le 3\sigma_2^2,$$
$$H_1: \quad \sigma_1^2 > 3\sigma_2^2.$$

Describe how to carry out an F test of these hypotheses.

9. Consider again the conditions of Exercise 7, but suppose now that it is desired to test the following hypotheses:

$$H_0: \quad \sigma_1^2 = \sigma_2^2,$$
$$H_1: \quad \sigma_1^2 \ne \sigma_2^2.$$

Suppose also that the statistic V is defined by Eq. (9.7.4), and it is desired to reject H_0 if either $V \le c_1$ or $V \ge c_2$, where the constants c_1 and c_2 are chosen so that when H_0 is true, $\Pr(V \le c_1) = \Pr(V \ge c_2) = 0.025$. Determine the values of c_1 and c_2 when $m = 16$ and $n = 10$, as in Exercise 7.

10. Suppose that a random sample consisting of 16 observations is available from the normal distribution for which both the mean μ_1 and the variance σ_1^2 are unknown, and an independent random sample consisting of 10 observations is available from the normal distribution for which both the mean μ_2 and the variance σ_2^2 are also unknown.

For each constant $r > 0$, construct a test of the following hypotheses at the level of significance 0.05:

$$H_0\colon \frac{\sigma_1^2}{\sigma_2^2} = r, \qquad H_1\colon \frac{\sigma_1^2}{\sigma_2^2} \neq r.$$

11. Consider again the conditions of Exercise 10. Use the results of that exercise to construct a confidence interval for σ_1^2/σ_2^2 with confidence coefficient 0.95.

12. Suppose that a random variable Y has the χ^2 distribution with m_0 degrees of freedom, and let c be a constant such that $\Pr(Y > c) = 0.05$. Explain why, in the table of 0.95 quantile of the F distribution, the entry for $m = m_0$ and $n = \infty$ will be equal to c/m_0.

13. The final column in the table of the 0.95 quantile of the F distribution contains values for which $m = \infty$. Explain how to derive the entries in this column from a table of the χ^2 distribution.

14. Consider again the conditions of Exercise 7. Find the power function of the F test when $\sigma_1^2 = 2\sigma_2^2$.

15. Prove Theorem 9.7.5. Also, compute the p-value for Example 9.7.4 using the formula in Eq. (9.7.8).

16. Let V be as defined in Eq. (9.7.4). We wish to determine the size α_0 likelihood ratio test of the hypotheses (9.7.7). Prove that the likelihood ratio test will reject H_0 if either $V \le c_1$ or $V \ge c_2$, where $\Pr(V \le c_1) + \Pr(V \ge c_2) = \alpha_0$ when $\sigma_1^2 = \sigma_2^2$.

17. Prove that the test found in Exercise 9 is *not* a likelihood ratio test.

18. Let δ be the two-sided F test that rejects H_0 in (9.7.3) when either $V \le c_1$ or $V \ge c_2$ with $c_1 < c_2$. Prove that the power function of δ is

$$\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|\delta) = G_{m-1,n-1}\left(\frac{\sigma_2^2}{\sigma_1^2}c_1\right) + 1 - G_{m-1,n-1}\left(\frac{\sigma_2^2}{\sigma_1^2}c_2\right).$$

19. Suppose that $X_1, \dots, X_{11}$ form a random sample from the normal distribution with unknown mean μ_1 and unknown variance σ_1^2. Suppose also that $Y_1, \dots, Y_{21}$ form an independent random sample from the normal distribution with unknown mean μ_2 and unknown variance σ_2^2. Suppose that we wish to test the hypotheses in Eq. (9.7.7). Let δ be the equal-tailed two-sided F test with level of significance $\alpha_0 = 0.5$.

a. Compute the power function of δ when $\sigma_1^2 = 1.01\sigma_2^2$.

b. Compute the power function of δ when $\sigma_1^2 = \sigma_2^2/1.01$.

c. Show that δ is not an unbiased test. (You will probably need computer software that computes the function $G_{m-1,n-1}$. And try to minimize the amount of rounding you do.)

★ 9.8 Bayes Test Procedures

Here we summarize how one tests hypotheses from the Bayesian perspective. The general idea is to choose the action (reject H_0 or not) that leads to the smaller posterior expected loss. We assume that the loss of making an incorrect decision is larger than the loss of making a correct decision. Many of the Bayes test procedures have the same forms as the tests we have already seen, but their interpretations are different.

Simple Null and Alternative Hypotheses

Example 9.8.1 Service Times in a Queue. In Example 9.2.1, a manager was trying to decide which of two joint distributions better describes customer service times. She was comparing the two joint p.d.f.'s f_1 and f_0 in Eqs. (9.2.1) and (9.2.2), respectively. Suppose that there are costs involved in making a bad choice. For example, if she chooses a joint distribution that models the service times as shorter than they really tend to be, there may be a cost due to customers becoming frustrated and taking their business elsewhere. On the other hand, if she chooses a joint distribution that models the service times as longer than they really tend to be, there may be a cost due to hiring additional unnecessary servers. How should the manager weigh these costs together with available evidence about how long she believes service times tend to be in order to choose between the two joint distributions? ◄

Consider a general problem in which the parameter space consists of two values $\Omega = \{\theta_0, \theta_1\}$. If $\theta = \theta_i$ (for $i = 0, 1$), let $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f_i(\boldsymbol{x})$. Suppose that it is desired to test the following simple hypotheses:

$$\begin{aligned} H_0&: \quad \theta = \theta_0, \\ H_1&: \quad \theta = \theta_1. \end{aligned} \tag{9.8.1}$$

We shall let d_0 denote the decision not to reject the hypothesis H_0 and let d_1 denote the decision to reject H_0. Also, we shall assume that the losses resulting from choosing an incorrect decision are as follows: If decision d_1 is chosen when H_0 is actually the true hypothesis (type I error), then the loss is w_0 units; if decision d_0 is chosen when H_1 is actually the true hypothesis (type II error), then the loss is w_1 units. If the decision d_0 is chosen when H_0 is the true hypothesis or if the decision d_1 is chosen when H_1 is the true hypothesis, then the correct decision has been made and the loss is 0. Thus, for $i = 0, 1$ and $j = 0, 1$, the loss $L(\theta_i, d_j)$ that occurs when θ_i is the true value of θ and the decision d_j is chosen is given by the following table:

	d_0	d_1
θ_0	0	w_0
θ_1	w_1	0

(9.8.2)

Next, suppose that the prior probability that H_0 is true is ξ_0, and the prior probability that H_1 is true is $\xi_1 = 1 - \xi_0$. Then the expected loss $r(\delta)$ of each test procedure δ will be

$$r(\delta) = \xi_0 E(\text{Loss } |\theta = \theta_0) + \xi_1 E(\text{Loss}|\theta = \theta_1). \tag{9.8.3}$$

If $\alpha(\delta)$ and $\beta(\delta)$ again denote the probabilities of the two types of errors for the procedure δ, and if the table of losses just given is used, it follows that

$$\begin{aligned} E(\text{Loss}|\theta = \theta_0) &= w_0 \Pr(\text{Choosing } d_1|\theta = \theta_0) = w_0\alpha(\delta), \\ E(\text{Loss}|\theta = \theta_1) &= w_1 \Pr(\text{Choosing } d_0|\theta = \theta_1) = w_1\beta(\delta). \end{aligned} \tag{9.8.4}$$

Hence,

$$r(\delta) = \xi_0 w_0 \alpha(\delta) + \xi_1 w_1 \beta(\delta). \tag{9.8.5}$$

A procedure δ for which this expected loss $r(\delta)$ is minimized is called a *Bayes test procedure*.

Since $r(\delta)$ is simply a linear combination of the form $a\alpha(\delta) + b\beta(\delta)$ with $a = \xi_0 w_0$ and $b = \xi_1 w_1$, a Bayes test procedure can immediately be determined from Theorem 9.2.1. Thus, a Bayes procedure will not reject H_0 whenever $\xi_0 w_0 f_0(\boldsymbol{x}) > \xi_1 w_1 f_1(\boldsymbol{x})$ and will reject H_0 whenever $\xi_0 w_0 f_0(\boldsymbol{x}) < \xi_1 w_1 f_1(\boldsymbol{x})$. We can either reject H_0 or not if $\xi_0 w_0 f_0(\boldsymbol{x}) = \xi_1 w_1 f_1(\boldsymbol{x})$. For simplicity, in the remainder of this section, we shall assume that H_0 is rejected whenever $\xi_0 w_0 f_0(\boldsymbol{x}) = \xi_1 w_1 f_1(\boldsymbol{x})$.

Note: Bayes Test Depends Only on the Ratio of Costs. Notice that choosing δ to minimize $r(\delta)$ in Eq. (9.8.5) is not affected if we multiply w_0 and w_1 by the same positive constant, such as $1/w_0$. That is, the Bayes test δ is also the test that minimizes

$$r^*(\delta) = \xi_0 \alpha(\delta) + \xi_1 \frac{w_1}{w_0} \beta(\delta).$$

So, a decision maker does not need to choose both of the two costs of error, but rather just the ratio of the two costs. One can think of choosing the ratio of costs as a replacement for specifying a level of significance when selecting a test procedure.

Example 9.8.2 Service Times in a Queue. Suppose that the manager believes that each of the two models for service times is equally likely before observing any data so that $\xi_0 = \xi_1 = 1/2$. The model with joint p.d.f. f_1 predicts both extremely large service times and extremely small service times to be more likely than does the model with joint p.d.f. f_0. Suppose that the cost of modeling extremely large service times as being less likely than they really are is the same as the cost of modeling extremely large service times to be more likely than they really are. The ratio of the cost of type II error w_1 to the cost of type I error w_0 is then $w_1/w_0 = 1$. The Bayes test is then to choose d_1 (reject H_0) if $f_0(\boldsymbol{x}) < f_1(\boldsymbol{x})$. This is equivalent to $f_1(\boldsymbol{x})/f_0(\boldsymbol{x}) > 1$. ◀

Tests Based on the Posterior Distribution

From the Bayesian viewpoint, it is more natural to base a test on the posterior distribution of θ rather than on the prior distribution and the probabilities of error as we did in the preceding discussion. Fortunately, the same test procedure arises regardless of how one derives it. For example, Exercise 5 in this section asks you to prove that the test derived by minimizing a linear combination of error probabilities is the same as what one would obtain by minimizing the posterior expected value of the loss. The same is true in general when the losses are bounded, but the proof is more difficult. For the remainder of this section, we shall take the more natural approach of trying to minimize the posterior expected value of the loss directly.

Return again to the general situation in which the null hypothesis is $H_0 : \theta \in \Omega_0$ and the alternative hypothesis is $H_1 : \theta \in \Omega_1$, where $\Omega_0 \cup \Omega_1$ is the entire parameter space. As we did above, we shall let d_0 denote the decision not to reject the null hypothesis H_0 and let d_1 denote the decision to reject H_0. As before, we shall assume that we incur a loss of w_0 by making decision d_1 when H_0 is actually true, and a loss of w_1 is incurred if we make decision d_0 when H_1 is true. (More realistic loss functions are available, but this simple type of loss will suffice for an introduction.) The loss function $L(\theta, d_i)$ can be summarized in the following table:

	d_0	d_1
If H_0 is true	0	w_0
If H_1 is true	w_1	0

(9.8.6)

We shall now take the approach outlined in Exercise 5. Suppose that $\xi(\theta|\boldsymbol{x})$ is the posterior p.d.f. for θ. Then the posterior expected loss $r(d_i|\boldsymbol{x})$ for choosing decision d_i $(i = 0, 1)$ is

$$r(d_i|\boldsymbol{x}) = \int L(\theta, d_i)\xi(\theta|\boldsymbol{x})\, d\theta.$$

We can write a simpler formula for this posterior expected loss for each of $i = 0, 1$:

$$r(d_0|\boldsymbol{x}) = \int_{\Omega_1} w_1\xi(\theta|\boldsymbol{x})\, d\theta = w_1[1 - \Pr(H_0 \text{ true}|\boldsymbol{x})],$$

$$r(d_1|\boldsymbol{x}) = \int_{\Omega_0} w_0\xi(\theta|\boldsymbol{x})\, d\theta = w_0 \Pr(H_0 \text{ true}|\boldsymbol{x}).$$

The Bayes test procedure is to choose the decision that has the smaller posterior expected loss, that is, choose d_0 if $r(d_0|\boldsymbol{x}) < r(d_1|\boldsymbol{x})$, choose d_1 if $r(d_0|\boldsymbol{x}) \geq r(d_1|\boldsymbol{x})$. Using the expressions above, it is easy to see that the inequality $r(d_0|\boldsymbol{x}) \geq r(d_1|\boldsymbol{x})$

(when to reject H_0) can be rewritten as

$$\Pr(H_0 \text{ true}|\boldsymbol{x}) \le \frac{w_1}{w_0 + w_1}, \tag{9.8.7}$$

just as in part (c) of Exercise 5.

The test procedure that rejects H_0 when (9.8.7) holds is the Bayes test in all situations in which the loss function is given by the table in (9.8.6). This result holds whether or not the distributions have monotone likelihood ratio, and it even applies when the alternative is two-sided or when the parameter is discrete rather than continuous. Furthermore, the Bayes test produces the same result if one were to switch the names of H_0 and H_1, as well as the losses w_0 and w_1 and the names of the decisions d_0 and d_1. (See Exercise 11 in this section.)

Despite the generality of (9.8.7), it is instructive to examine what the procedure looks like in special cases that we have already encountered.

One-Sided Hypotheses

Suppose that the family of distributions has a monotone likelihood ratio and that the hypotheses are

$$\begin{aligned} H_0&: \quad \theta \le \theta_0, \\ H_1&: \quad \theta > \theta_0. \end{aligned} \tag{9.8.8}$$

We shall prove next that the Bayes procedure that rejects H_0 when (9.8.7) holds is a one-sided test as in Theorem 9.3.1.

Theorem 9.8.1 Suppose that $f_n(\boldsymbol{x}|\theta)$ has a monotone likelihood ratio in the statistic $T = r(\boldsymbol{X})$. Let the hypotheses be as in Eq. (9.8.8), and assume that the loss function is of the form

	d_0	d_1
$\theta \le \theta_0$	0	w_0
$\theta > \theta_0$	w_1	0

where $w_0, w_1 > 0$ are constants. Then a test procedure that minimizes the posterior expected loss is to reject H_0 when $T \ge c$ for some constant c (possibly infinite).

Proof According to Bayes' theorem for parameters and samples, (7.2.7), the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ can be expressed as

$$\xi(\theta|\boldsymbol{x}) = \frac{f_n(\boldsymbol{x}|\theta)\xi(\theta)}{\int_\Omega f_n(\boldsymbol{x}|\psi)\xi(\psi)\,d\psi}.$$

The ratio of the posterior expected loss from making decision d_0 to the posterior expected loss from making decision d_1 after observing $\boldsymbol{X} = \boldsymbol{x}$ is

$$\ell(\boldsymbol{x}) = \frac{\int_{\theta_0}^{\infty} w_1\xi(\theta|\boldsymbol{x})\,d\theta}{\int_{-\infty}^{\theta_0} w_0\xi(\psi|\boldsymbol{x})\,d\psi} = \frac{w_1\int_{\theta_0}^{\infty} f_n(\boldsymbol{x}|\theta)\xi(\theta)\,d\theta}{w_0\int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}|\psi)\xi(\psi)\,d\psi}. \tag{9.8.9}$$

What we need to prove is that $\ell(\boldsymbol{x}) \ge 1$ is equivalent to $T \ge c$. It suffices to show that $\ell(\boldsymbol{x})$ is a nondecreasing function in $T = r(\boldsymbol{x})$. Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be two possible observations with the property that $r(\boldsymbol{x}_1) \le r(\boldsymbol{x}_2)$. We want to prove that $\ell(\boldsymbol{x}_1) \le \ell(\boldsymbol{x}_2)$.

We can write

$$\ell(\boldsymbol{x}_1) - \ell(\boldsymbol{x}_2) = \frac{w_1 \int_{\theta_0}^{\infty} f_n(\boldsymbol{x}_1|\theta)\xi(\theta)\,d\theta}{w_0 \int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}_1|\psi)\xi(\psi)\,d\psi} - \frac{w_1 \int_{\theta_0}^{\infty} f_n(\boldsymbol{x}_2|\theta)\xi(\theta)\,d\theta}{w_0 \int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}_2|\psi)\xi(\psi)\,d\psi}. \tag{9.8.10}$$

We can put the two fractions on the right side of Eq. (9.8.10) over the common denominator $w_0^2 \int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}_2|\psi)\xi(\psi)\,d\psi \int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}_1|\psi)\xi(\psi)\,d\psi$. The numerator of the resulting fraction is $w_0 w_1$ times

$$\begin{aligned} &\int_{\theta_0}^{\infty} f_n(\boldsymbol{x}_1|\theta)\xi(\theta)\,d\theta \int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}_2|\psi)\xi(\psi)\,d\psi \\ &- \int_{\theta_0}^{\infty} f_n(\boldsymbol{x}_2|\theta)\xi(\theta)\,d\theta \int_{-\infty}^{\theta_0} f_n(\boldsymbol{x}_1|\psi)\xi(\psi)\,d\psi. \end{aligned} \tag{9.8.11}$$

We only need to show that (9.8.11) is at most 0. The difference in (9.8.11) can be written as the double integral

$$\int_{\theta_0}^{\infty} \int_{-\infty}^{\theta_0} \xi(\theta)\xi(\psi)[f_n(\boldsymbol{x}_1|\theta)f_n(\boldsymbol{x}_2|\psi) - f_n(\boldsymbol{x}_2|\theta)f_n(\boldsymbol{x}_1|\psi)]\,d\psi\,d\theta. \tag{9.8.12}$$

Notice that for all θ and ψ in this double integral, $\theta \ge \theta_0 \ge \psi$. Since $r(\boldsymbol{x}_1) \le r(\boldsymbol{x}_2)$, monotone likelihood ratio implies that

$$\frac{f_n(\boldsymbol{x}_1|\theta)}{f_n(\boldsymbol{x}_1|\psi)} - \frac{f_n(\boldsymbol{x}_2|\theta)}{f_n(\boldsymbol{x}_2|\psi)} \le 0.$$

If one multiplies both sides of this last expression by the product of the two denominators, the result is

$$f_n(\boldsymbol{x}_1|\theta)f_n(\boldsymbol{x}_2|\psi) - f_n(\boldsymbol{x}_2|\theta)f_n(\boldsymbol{x}_1|\psi) \le 0. \tag{9.8.13}$$

Notice that the left side of Eq. (9.8.13) appears inside the square brackets in the integrand of (9.8.12). Since this is nonpositive, it implies that (9.8.12) is at most 0, and so (9.8.11) is at most 0. ■

Example 9.8.3

Calorie Counts on Food Labels. In Example 7.3.10 on page 400, we were interested in the percentage differences between the observed and advertised calorie counts for nationally prepared foods. We modeled the differences $X_1, \ldots, X_{20}$ as normal random variables with mean θ and variance 100. The prior for θ was a normal distribution with mean 0 and variance 60. The family of normal distributions has a monotone likelihood ratio in the statistic $\overline{X}_{20} = \frac{1}{20}\sum_{i=1}^{20} X_i$. The posterior distribution of θ is the normal distribution with mean

$$\mu_1 = \frac{100 \times 0 + 20 \times 60 \times \overline{X}_{20}}{100 + 20 \times 60} = 0.923\overline{X}_{20}$$

and variance $v_1^2 = 4.62$. Suppose that we wish to test the null hypothesis $H_0 : \theta \le 0$ versus the alternative $H_1 : \theta > 0$. The posterior probability that H_0 is true is

$$\Pr(\theta \le 0|\overline{X}_{20}) = \Phi\left(\frac{0 - \mu_1}{v_1}\right) = \Phi(-0.429\overline{X}_{20}).$$

The Bayes test will reject H_0 if this probability is at most $w_1/(w_0 + w_1)$. Since Φ is a strictly increasing function, $\Phi(-0.429\overline{X}_{20}) \le w_1/(w_0 + w_1)$ if and only if $\overline{X}_{20} \ge -\Phi^{-1}(w_1/(w_0 + w_1))/0.492$. This is in the form of a one-sided test. ◀

Two-Sided Alternatives

On page 571, we argued that the hypotheses

$$\begin{aligned} H_0&: \quad \theta = \theta_0, \\ H_1&: \quad \theta \neq \theta_0 \end{aligned} \tag{9.8.14}$$

might be a useful surrogate for the null hypothesis that θ is close to θ_0 against the alternative that it is not close. If the prior distribution of θ is continuous, then the posterior distribution will usually be continuous as well. In such cases, the posterior probability that H_0 is true will be 0, and H_0 would be rejected without having to refer to the data. If one believed that $\theta = \theta_0$ with positive probability, one should use a prior distribution that is not continuous, but we shall not take that approach here. (See a more advanced text, such as Schervish, 1995, section 4.2, for treatment of that approach.) Instead, we can calculate the posterior probability that θ is close to θ_0. If this probability is too small, we can reject the null hypothesis that θ is close to θ_0. To be specific, let $d > 0$, and consider the hypotheses

$$\begin{aligned} H_0&: \quad |\theta - \theta_0| \le d, \\ H_1&: \quad |\theta - \theta_0| > d. \end{aligned} \tag{9.8.15}$$

Many experimenters might choose to test the hypotheses in (9.8.14) rather than those in (9.8.15) because they are not ready to specify a particular value of d. In such cases, one could calculate the posterior probability of $|\theta - \theta_0| \le d$ for all d and draw a little plot.

Example 9.8.4 **Calorie Counts on Food Labels.** Suppose that we wish to test the hypotheses (9.8.15) with $\theta_0 = 0$ in the situation described in Example 9.8.3. In Example 7.3.10, we found that the posterior distribution of θ was the normal distribution with mean 0.1154 and variance 4.62. We can easily calculate

$$\Pr(|\theta - 0| \le d|\boldsymbol{x}) = \Pr(-d \le \theta \le d|\boldsymbol{x}) = \Phi\left(\frac{d - 0.1154}{4.62^{1/2}}\right) - \Phi\left(\frac{-d - 0.1154}{4.62^{1/2}}\right),$$

for every value of d that we want. Figure 9.16 shows a plot of the posterior probability that $|\theta|$ is at most d for all values of d between 0 and 5. In particular, we see that $\Pr(|\theta| \le 5|\boldsymbol{x})$ is very close to 1. If 5 percent is considered a small discrepancy, then we can be pretty sure that $|\theta|$ is small. On the other hand, $\Pr(|\theta| \ge 1|\boldsymbol{x})$ is greater than 0.6. If 1 percent is considered large, then there is a substantial chance that $|\theta|$ is large. ◀

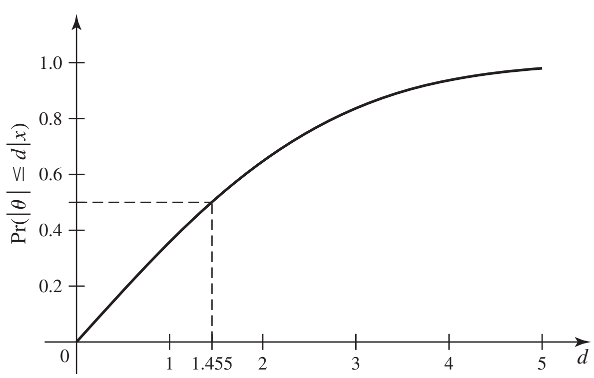


Figure 9.16 Plot of $\Pr(|\theta| \le d|\boldsymbol{x})$ against d for Example 9.8.4. The dotted lines indicate that the median of the posterior distribution of $|\theta|$ is 1.455.

Note: What Counts as a Meaningful Difference?. The method illustrated in Example 9.8.4 raises a useful point. In order to complete the test procedure, we need to decide what counts as a meaningful difference between θ and θ_0. Otherwise, we cannot say whether or not the probability is large that a meaningful difference exists. Forcing experimenters to think about what counts as a meaningful difference is a good idea. Testing the hypotheses (9.8.14) at a fixed level, such as 0.05, does not require anyone to think about what counts as a meaningful difference. Indeed, if an experimenter did bother to decide what counted as a meaningful difference, it is not clear how to make use of that information in choosing a significance level at which to test the hypotheses in (9.8.14).

Testing the Mean of a Normal Distribution with Unknown Variance

In Sec. 8.6, we considered the case in which a random sample is drawn from a normal distribution with unknown mean and variance. We introduced a family of conjugate prior distributions and found that the posterior distribution of a linear function of the mean μ is a t distribution. If we wish to test the null hypothesis that μ lies in an interval using (9.8.7) as the condition for rejecting the null hypothesis, then we only need a table or computer program to calculate the c.d.f. of an arbitrary t distribution. Most statistical software packages allow calculation of the c.d.f. and the quantile function of an arbitrary t distribution, and hence we can perform Bayes tests of null hypotheses of the form $\mu \le \mu_0$, $\mu \ge \mu_0$, or $d_1 \le \mu \le d_2$.

Example 9.8.5 Pesticide Residue on Celery. Sharpe and Van Middelem (1955) describe an experiment in which $n = 77$ samples of parathion residue were measured on celery after the vegetable had been taken from fields sprayed with parathion. Figure 9.17 shows a histogram of the observations. (Each concentration Z in parts per million was transformed to $X = 100(Z - 0.7)$ for ease of recording.) Suppose that we model the X values as normal with mean μ and variance σ^2. We will use an improper prior for μ and σ^2. The sample average is $\overline{x}_n = 50.23$, and

$$s_n^2 = \sum_{i=1}^{77} (x_i - \overline{x}_{77})^2 = 34106.$$

As we saw in Eq. (8.6.21), this means that the posterior distribution of

$$\frac{n^{1/2}(\mu - \overline{x}_n)}{(s_n^2/(n-1))^{1/2}} = \frac{77^{1/2}(\mu - 50.23)}{(34106/76)^{1/2}} = 0.4142\mu - 20.81$$

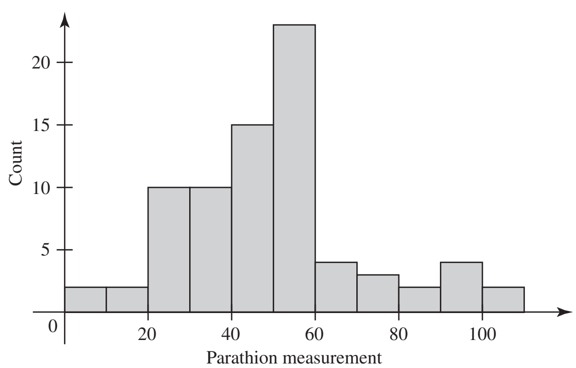


Figure 9.17 Histogram of parathion measurements on 77 celery samples.

is the t distribution with 76 degrees of freedom. Suppose that we are interested in testing the null hypothesis $H_0: \mu \geq 55$ against the alternative $H_1: \mu < 55$. Suppose that our losses are described by (9.8.6). Then we should reject H_0 if its posterior probability is at most $\alpha_0 = w_1/(w_0 + w_1)$. If we let T_{n-1} stand for the c.d.f. of the t distribution with $n-1$ degrees of freedom, we can write this probability as

$$\begin{aligned}\Pr(\mu \geq 55|\boldsymbol{x}) &= \Pr\left(\frac{n^{1/2}(\mu - \overline{x}_n)}{(s_n^2/(n-1))^{1/2}} \geq \frac{n^{1/2}(55 - \overline{x}_n)}{(s_n^2/(n-1))^{1/2}} \middle| \boldsymbol{x} \right) \\ &= 1 - T_{n-1}\left(\frac{n^{1/2}(55 - \overline{x}_n)}{(s_n^2/(n-1))^{1/2}} \right). \end{aligned} \tag{9.8.16}$$

Simple manipulation shows that this last probability is at most α_0 if and only if $U \leq T_{n-1}^{-1}(1 - \alpha_0)$, where U is the random variable in Eq. (9.5.2) that was used to define the t test. Indeed, the level α_0 t test of H_0 versus H_1 is precisely to reject H_0 if $U \leq T_{n-1}^{-1}(1 - \alpha_0)$. For the data in this example, the probability in Eq. (9.8.16) is $1 - T_{76}(1.974) = 0.026$. ◀

Note: Look at Your Data. The histogram in Fig. 9.17 has a strange feature. Can you specify what it is? If you take a course in data analysis, you will probably learn some methods for dealing with data having features like this.

Note: Bayes Tests for One-Sided Nulls with Improper Priors Are *t* Tests. In Example 9.8.5, we saw that the Bayes test for one-sided hypotheses was the level α_0 t test for the same hypotheses where $\alpha_0 = w_1/(w_0 + w_1)$. This holds in general for normal data with improper priors. It also follows that the p-values in these cases must be the same as the posterior probabilities that the null hypotheses are true. (See Exercise 7 in this section.)

Comparing the Means of Two Normal Distributions

Next, consider the case in which we shall observe two independent normal random samples with common variance σ^2: $X_1, \ldots, X_m$ with mean μ_1 and $Y_1, \ldots, Y_n$ with mean μ_2. In order to use the Bayesian approach, we need the posterior distribution of $\mu_1 - \mu_2$. We could introduce a family of conjugate prior distributions for the three parameters μ_1, μ_2, and $\tau = 1/\sigma^2$, and then proceed as we did in Sec. 8.6. For simplicity, we shall only handle the case of improper priors in this section, although there are proper conjugate priors that will lead to more general results. The usual improper prior for each parameter μ_1 and μ_2 is the constant function 1, and the usual improper prior for τ is $1/\tau$ for $\tau > 0$. If we combine these as if the parameters were independent, the improper prior p.d.f. would be $\xi(\mu_1, \mu_2, \tau) = 1/\tau$ for $\tau > 0$. We can now find the posterior joint distribution of the parameters.

Theorem 9.8.2 Suppose that $X_1, \ldots, X_m$ form a random sample from a normal distribution with mean μ_1 and precision τ while $Y_1, \ldots, Y_n$ form a random sample from a normal distribution with mean μ_2 and precision τ. Suppose that the parameters have the improper prior with "p.d.f." $\xi(\mu_1, \mu_2, \tau) = 1/\tau$ for $\tau > 0$. The posterior distribution of

$$(m+n-2)^{1/2} \frac{\mu_1 - \mu_2 - (\overline{x}_m - \overline{y}_n)}{\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2} (s_x^2 + s_y^2)^{1/2}} \tag{9.8.17}$$

is the t distribution with $m+n-2$ degrees of freedom, where s_x^2 and s_y^2 are the observed values of S_X^2 and S_Y^2, respectively. ■

The proof of Theorem 9.8.2 is left as Exercise 8 because it is very similar to results proven in Sec. 8.6.

For testing the hypotheses

$$\begin{aligned} H_0&: \quad \mu_1 - \mu_2 \le 0, \\ H_1&: \quad \mu_1 - \mu_2 > 0, \end{aligned}$$

we need the posterior probability that $\mu_1 - \mu_2 \le 0$, which is easily obtained from the posterior distribution. Using the same idea as in Eq. (9.8.16), we can write $\Pr(\mu_1 - \mu_2 \le 0|\boldsymbol{x}, \boldsymbol{y})$ as the probability that the random variable in (9.8.17) is at most $-u$, where u is the observed value of the random variable U in Eq. (9.6.3). It follows that

$$\Pr(\mu_1 - \mu_2 \le 0|\boldsymbol{x}, \boldsymbol{y}) = T_{m+n-2}(-u),$$

where T_{m+n-2} is the c.d.f. of the t distribution with $m+n-2$ degrees of freedom. Hence, the posterior probability that H_0 is true is less than $w_1/(w_0+w_1)$ if and only if

$$T_{m+n-2}(-u) < \frac{w_1}{w_0+w_1}.$$

This, in turn is true if and only if

$$-u < T^{-1}_{m+n-2}\left(\frac{w_1}{w_0+w_1}\right).$$

This is true if and only if

$$u > T^{-1}_{m+n-2}\left(1-\frac{w_1}{w_0+w_1}\right). \tag{9.8.18}$$

If $\alpha_0 = w_1/(w_0+w_1)$, then the Bayes test procedure that rejects H_0 when Eq. (9.8.18) occurs is the same as the level α_0 two-sample t test derived in Sec. 9.6. Put another way, the one-sided level α_0 two-sample t test rejects the null hypothesis H_0 if and only if the posterior probability that H_0 is true (based on the improper prior) is at most α_0. It follows from Exercise 7 that the posterior probability of the null hypothesis being true must equal the p-value in this case.

Example 9.8.6 Roman Pottery in Britain. In Example 9.6.3, we observed 14 samples of Roman pottery from Llanederyn in Great Britain and another five samples from Ashley Rails, and we were interested in whether the mean aluminum oxide percentage in Llanederyn μ_1 was larger than that in Ashley Rails μ_2. We tested $H_0: \mu_1 \ge \mu_2$ against $H_1: \mu_1 < \mu_2$ and found that the p-value was 4×10^{-6}. If we had used an improper prior for the parameters, then $\Pr(\mu_1 \ge \mu_2|\boldsymbol{x}) = 4 \times 10^{-6}$. ◄

Two-Sided Alternatives with Unknown Variance To test the hypothesis that the mean μ of a normal distribution is close to μ_0, we could specify a specific value d and test

$$\begin{aligned} H_0&: \quad |\mu - \mu_0| \le d, \\ H_1&: \quad |\mu - \mu_0| > d. \end{aligned}$$

If we do not feel comfortable selecting a single value of d to represent "close," we could compute $\Pr(|\mu - \mu_0| \le d|\boldsymbol{x})$ for all d and draw a plot as we did in Example 9.8.4.

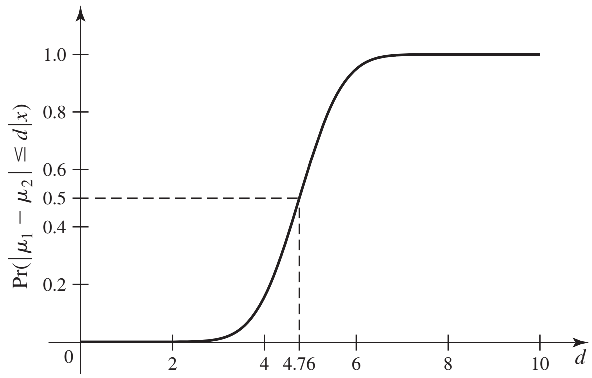


Figure 9.18 Plot of $\Pr(|\mu_1 - \mu_2| \le d|\boldsymbol{x})$ against d. The dotted lines indicate that the median of the posterior distribution of $|\mu_1 - \mu_2|$ is 4.76.

The case of testing that two means are close together can be dealt with in the same way.

Example 9.8.7 Roman Pottery in Britain. In Example 9.8.6, we tested one-sided hypotheses about the difference in aluminum oxide contents in samples of pottery from two sites in Great Britain. Unless we are specifically looking for a difference in a particular direction, it might make more sense to test hypotheses of the form

$$\begin{aligned} H_0&: \quad |\mu_1 - \mu_2| \le d, \\ H_1&: \quad |\mu_1 - \mu_2| > d, \end{aligned} \tag{9.8.19}$$

where d is some critical difference that is worth detecting. As we did in Example 9.8.4, we can draw a plot that allows us to test all hypotheses of the form (9.8.19) simultaneously. We just plot $\Pr(|\mu_1 - \mu_2| \le d|\boldsymbol{x})$ against d. The posterior distribution of $\mu_1 - \mu_2$ was found in Eq. (9.8.17), using the improper prior. In this case, the following random variable has the t distribution with 17 degrees of freedom:

$$\begin{aligned} &(m+n-2)^{1/2}\frac{\mu_1 - \mu_2 - (\overline{x}_m - \overline{y}_n)}{\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}(s_x^2 + s_y^2)^{1/2}} \\ &\quad = 17^{1/2}\frac{\mu_1 - \mu_2 - (12.56 - 17.32)}{\left(\frac{1}{14}+\frac{1}{5}\right)^{1/2}(24.65 + 11.01)^{1/2}} = 1.33(\mu_1 - \mu_2 + 4.76), \end{aligned}$$

where the data summaries come from Example 9.6.3. It follows that

$$\begin{aligned} &\Pr(|\mu_1 - \mu_2| \le d|\boldsymbol{x}) \\ &\quad = \Pr(1.33(-d + 4.76) \le 1.33(\mu_1 - \mu_2 + 4.76) \le 1.33(d + 4.76)|\boldsymbol{x}) \\ &\quad = T_{17}(1.33(d + 4.76)) - T_{17}(1.33(-d + 4.76)), \end{aligned}$$

where T_{17} is the c.d.f. of the t distribution with 17 degrees of freedom. Figure 9.18 is the plot of this posterior probability against d. ◄

Comparing the Variances of Two Normal Distributions

In order to test hypotheses concerning the variances of two normal distributions, we can make use of the posterior distribution of the ratio of the two variances. Suppose that $X_1, \dots, X_m$ is a random sample from the normal distribution with mean μ_1 and variance σ_1^2, and $Y_1, \dots, Y_n$ is a random sample from the normal distribution with mean μ_2 and variance σ_2^2. If we model the X data and associated parameters

as independent of the Y data and associated parameters, then we can perform two separate analyses just like the one in Sec. 8.6. In particular, we let $\tau_i = 1/\sigma_i^2$ for $i = 1, 2$, and the joint posterior distribution will have (μ_1, τ_1) independent of (μ_2, τ_2) and each pair will have a normal-gamma distribution just as in Sec. 8.6. For convenience, we shall only do the remaining calculations using improper priors. With improper priors, the posterior distribution of τ_1 is the gamma distribution with parameters $(m-1)/2$ and $s_x^2/2$, where s_x^2 is defined in Theorem 9.8.2. We also showed in Sec. 8.6 (using Exercise 1 in Sec. 5.7) that $\tau_1 s_x^2$ has the χ^2 distribution with $m-1$ degrees of freedom. Similarly, $\tau_2 s_y^2$ has the χ^2 distribution with $n-1$ degrees of freedom. Since $\tau_1 s_x^2/(m-1)$ and $\tau_2 s_y^2/(n-1)$ are independent, their ratio has the F distribution with $m-1$ and $n-1$ degrees of freedom. That is, the posterior distribution of

$$\frac{\tau_1 s_x^2/(m-1)}{\tau_2 s_y^2/(n-1)} = \frac{s_x^2/[(m-1)\sigma_1^2]}{s_y^2/[(n-1)\sigma_2^2]} \tag{9.8.20}$$

is the F distribution with $m-1$ and $n-1$ degrees of freedom. Notice that the expression on the right side of Eq. (9.8.20) is the same as the random variable V^* in Eq. (9.7.5). This is another case in which the sampling distribution of a random variable is the same as its posterior distribution. It will then follow that level α_0 tests of one-sided hypotheses about σ_1^2/σ_2^2 based on the sampling distribution of V^* will be the same as Bayes tests of the form (9.8.7) so long as $\alpha_0 = w_1/(w_0 + w_1)$. The reader can prove this in Exercise 9.

Summary

From a Bayesian perspective, one chooses a test procedure by minimizing the posterior expected loss. When the loss has the simple form of (9.8.6), then the Bayes test procedure is to reject H_0 when its posterior probability is at most $w_1/(w_0 + w_1)$. In many one-sided cases, with improper priors, this procedure turns out to be the same as the most commonly used level $\alpha_0 = w_1/(w_0 + w_1)$ test. In two-sided cases, as an alternative to testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, one can draw a plot of $\Pr(|\theta - \theta_0| \le d|\boldsymbol{x})$ against d. One then needs to decide which values of d count as meaningful differences.

Exercises

1. Suppose that a certain industrial process can be either in control or out of control, and that at any specified time the prior probability that it will be in control is 0.9, and the prior probability that it will be out of control is 0.1. A single observation X of the output of the process is to be taken, and it must be decided immediately whether the process is in control or out of control. If the process is in control, then X will have the normal distribution with mean 50 and variance 1.If the process is out of control, then X will have the normal distribution with mean 52 and variance 1.

If it is decided that the process is out of control when in fact it is in control, then the loss from unnecessarily stopping the process will be \$1000. If it is decided that the process is in control when in fact it is out of control, then the loss from continuing the process will be \$18,000. If a correct decision is made, then the loss will be 0. It is desired to find a test procedure for which the expected loss will be a minimum. For what values of X should it be decided that the process is out of control?

2. A single observation X is to be taken from a continuous distribution for which the p.d.f. is either f_0 or f_1, where

$$f_0(x) = \begin{cases} 1 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_1(x) = \begin{cases} 4x^3 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

On the basis of the observation X, it must be decided whether f_0 or f_1 is the correct p.d.f. Suppose that the prior probability that f_0 is correct is 2/3 and the prior probability that f_1 is correct is 1/3. Suppose also that the loss from choosing the correct decision is 0, the loss from deciding that f_1 is correct when in fact f_0 is correct is 1 unit, and the loss from deciding that f_0 is correct when in fact f_1 is correct is 4 units. If the expected loss is to be minimized, for what values of X should it be decided that f_0 is correct?

3. Suppose that a failure in a certain electronic system can occur because of either a minor or a major defect. Suppose also that 80 percent of the failures are caused by minor defects, and 20 percent of the failures are caused by major defects. When a failure occurs, n independent soundings $X_1, \dots, X_n$ are made on the system. If the failure was caused by a minor defect, these soundings form a random sample from the Poisson distribution with mean 3. If the failure was caused by a major defect, these soundings form a random sample from a Poisson distribution for which the mean is 7. The cost of deciding that the failure was caused by a major defect when it was actually caused by a minor defect is \$400. The cost of deciding that the failure was caused by a minor defect when it was actually caused by a major defect is \$2500. The cost of choosing a correct decision is 0. For a given set of observed values of $X_1, \dots, X_n$, which decision minimizes the expected cost?

4. Suppose that the proportion p of defective items in a large manufactured lot is unknown, and it is desired to test the following simple hypotheses:

$$H_0:\quad p = 0.3,$$
$$H_1:\quad p = 0.4.$$

Suppose that the prior probability that $p = 0.3$ is 1/4, and the prior probability that $p = 0.4$ is 3/4; also suppose that the loss from choosing an incorrect decision is 1 unit, and the loss from choosing a correct decision is 0. Suppose that a random sample of n items is selected from the lot. Show that the Bayes test procedure is to reject H_0 if and only if the proportion of defective items in the sample is greater than

$$\frac{\log\left(\frac{7}{6}\right) + \frac{1}{n}\log\left(\frac{1}{3}\right)}{\log\left(\frac{14}{9}\right)}.$$

5. Suppose that we wish to test the hypotheses (9.8.1). Let the loss function have the form of (9.8.2).

a. Prove that the posterior probability of $\theta = \theta_0$ is $\xi_0 f_0(\boldsymbol{x})/[\xi_0 f_0(\boldsymbol{x}) + \xi_1 f_1(\boldsymbol{x})]$.

b. Prove that a test that minimizes $r(\delta)$ also minimizes the posterior expected value of the loss given $\boldsymbol{X} = \boldsymbol{x}$ for all $\boldsymbol{x}$.

c. Prove that the following test is one of the tests described in part (b): "reject H_0 if $\Pr(H_0 \text{ true}|\boldsymbol{x}) \le w_1/(w_0 + w_1)$."

6. Prove that the conclusion of Theorem 9.8.1 still holds when the loss function is given by

	d_0	d_1
$\theta \le \theta_0$	0	$w_0(\theta)$
$\theta > \theta_0$	$w_1(\theta)$	0

for arbitrary positive functions $w_0(\theta)$ and $w_1(\theta)$. *Hint:* Replicate the proof of Theorem 9.8.1, but replace the constants w_0 and w_1 by the functions above and keep them inside of the integrals instead of factoring them out.

7. Suppose that we have a situation in which the Bayes test that rejects H_0 when $\Pr(H_0 \text{ true }|\boldsymbol{x}) \le \alpha_0$ is the same as the level α_0 test of H_0 for all α_0. (Example 9.8.5 has this property, but so do many other situations.) Prove that the p-value equals the posterior probability that H_0 is true.

8. In this exercise you will prove Theorem 9.8.2.

a. Prove that the joint p.d.f. of the data given the parameters μ_1, μ_2, and τ can be written as a constant times

$$\tau^{(m+n)/2}\exp\Big(-0.5m\tau(\mu_1 - \overline{x}_m)^2 - 0.5n\tau(\mu_2 - \overline{y}_n)^2 - 0.5(s_x^2 + s_y^2)\tau\Big).$$

b. Multiply the prior p.d.f. times the p.d.f. in part (a). Bayes' theorem for random variables says that the result is proportional (as a function of the parameters) to the posterior p.d.f.

i. Show that the posterior p.d.f., as a function of μ_1 for fixed μ_2 and τ, is the p.d.f. of the normal distribution with mean $\overline{x}_m$ and variance $(m\tau)^{-1}$.

ii. Show that the posterior p.d.f., as a function of μ_2 for fixed μ_1 and τ, is the p.d.f. of the normal distribution with mean $\overline{y}_n$ and variance $(n\tau)^{-1}$.

iii. Show that, conditional on τ, μ_1 and μ_2 are independent with the two normal distributions found above.

iv. Show that the marginal posterior distribution of τ is the gamma distribution with parameters $(m + n - 2)/2$ and $(s_x^2 + s_y^2)/2$.

c. Show that the conditional distribution of

$$Z = \tau^{1/2}\frac{\mu_1 - \mu_2 - (\overline{x}_m - \overline{y}_n)}{\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}}$$

given τ is a standard normal distribution and hence Z is independent of τ.

d. Show that the distribution of $W = (s_x^2 + s_y^2)\tau$ is the gamma distribution with parameters $(m + n - 2)/2$ and $1/2$, which is the same as the χ^2 distribution with $m + n - 2$ degrees of freedom.

e. Prove that $Z/(W/(m + n - 2))^{1/2}$ has the t distribution with $m + n - 2$ degrees of freedom and that it equals the expression in Eq. (9.8.17).

9. Suppose that $X_1, \dots, X_m$ form a random sample from the normal distribution with mean μ_1 and variance σ_1^2, and $Y_1, \dots, Y_n$ form a random sample from the normal distribution with mean μ_2 and variance σ_2^2. Suppose that we use the usual improper prior and that we wish to test the hypotheses

$$\begin{aligned} H_0:&\quad \sigma_1^2 \le \sigma_2^2,\\ H_1:&\quad \sigma_1^2 > \sigma_2^2. \end{aligned}$$

a. Prove that the level α_0 F test is the same as the test in (9.8.7) when $\alpha_0 = w_1/(w_0 + w_1)$.

b. Prove that the p-value for the F test is the posterior probability that H_0 is true.

10. Consider again the situation in Example 9.6.2. Let μ_1 be the mean of log-rainfall from seeded clouds, and let μ_2 be the mean of log-rainfall from unseeded clouds. Use the improper prior for the parameters.

a. Find the posterior distribution of $\mu_1 - \mu_2$.

b. Draw a graph of the posterior probability that $|\mu_1 - \mu_2| \le d$ as a function of d.

11. Let θ be a general parameter taking values in a parameter space Ω. Let $\Omega' \bigcup \Omega'' = \Omega$ be a partition of Ω into two disjoint sets Ω' and Ω''. We want to choose between two decisions: d' says that $\theta \in \Omega'$, and d'' says that $\theta \in \Omega''$. We have the following loss function:

	d'	d''
If $\theta \in \Omega'$	0	w'
If $\theta \in \Omega''$	w''	0

We have two choices for expressing this decision problem as a hypothesis-testing problem. One choice would be to define $H_0 : \theta \in \Omega'$ and $H_1 : \theta \in \Omega''$. The other choice would be to define $H_0 : \theta \in \Omega''$ and $H_1 : \theta \in \Omega'$. In this problem, we show that the Bayes test makes the same decision regardless of which hypothesis we call the null and which we call the alternative.

a. For each choice, say how we would define each of the following in order to make this problem fit the hypothesis-testing framework described in this section: w_0, w_1, d_0, d_1, Ω_0, and Ω_1.

b. Now suppose that we can observe data $\boldsymbol{X} = \boldsymbol{x}$ and compute the posterior distribution of θ, $\xi(\theta|\boldsymbol{x})$. Show that, for each of the two setups constructed in the previous part, the Bayes test chooses the same decision d' or d''. That is, observing $\boldsymbol{x}$ leads to choosing d' in the first setup if and only if observing $\boldsymbol{x}$ leads to choosing d' in the second setup. Similarly, observing $\boldsymbol{x}$ leads to choosing d'' in the first setup if and only if observing $\boldsymbol{x}$ leads to choosing d'' in the second setup.

⋆ 9.9 Foundational Issues

We discuss the relationship between significance level and sample size. We also distinguish between results that are significant in the statistical sense and those that are significant in a practical sense.

The Relationship between Level of Significance and Sample Size

In many statistical applications, it has become standard practice for an experimenter to specify a level of significance α_0, and then to find a test procedure with a large power function on the alternative hypothesis among all procedures whose size $\alpha(\delta) \le \alpha_0$. Alternatively, the experimenter will compute a p-value and report whether or not it was less than α_0. For the case of testing simple null and alternative hypotheses, the Nayman-Pearson lemma explicitly describes how to construct such a procedure. Furthermore, it has become traditional in many applications to choose the level of significance α_0 to be 0.10, 0.05, or 0.01. The selected level depends on how serious the consequences of an error of type I are judged to be. The value of α_0 most commonly used is 0.05. If the consequences of an error of type I are judged to be relatively mild in a particular problem, the experimenter may choose α_0 to be 0.10. On the other

hand, if these consequences are judged to be especially serious, the experimenter may choose α_0 to be 0.01.

Because these values of α_0 have become established in statistical practice, the choice of $\alpha_0 = 0.01$ is sometimes made by an experimenter who wishes to use a cautious test procedure, or one that will not reject H_0 unless the sample data provide strong evidence that H_0 is not true. We shall now show, however, that when the sample size n is large, the choice of $\alpha_0 = 0.01$ can actually lead to a test procedure that will reject H_0 for certain samples that, in fact, provide stronger evidence for H_0 than they do for H_1.

To illustrate this property, suppose, as in Example 9.2.5, that a random sample is taken from the normal distribution with unknown mean θ and known variance 1, and that the hypotheses to be tested are

$$\begin{aligned} H_0&: \quad \theta = 0, \\ H_1&: \quad \theta = 1. \end{aligned}$$

It follows from the discussion in Example 9.2.5 that, among all test procedures for which $\alpha(\delta) \le 0.01$, the probability of type II error $\beta(\delta)$ will be a minimum for the procedure δ^* that rejects H_0 when $\overline{X}_n \ge k'$, where k' is chosen so that $\Pr(\overline{X}_n \ge k'|\theta = 0) = 0.01$. When $\theta = 0$, the random variable $\overline{X}_n$ has the normal distribution with mean 0 and variance $1/n$. Therefore, it can be found from a table of the standard normal distribution that $k' = 2.326n^{-1/2}$.

Furthermore, it follows from Eq. (9.2.12) that this test procedure δ^* is equivalent to rejecting H_0 when $f_1(\boldsymbol{x})/f_0(\boldsymbol{x}) \ge k$, where $k = \exp(2.326n^{1/2} - 0.5n)$. The probability of an error of type I will be $\alpha(\delta^*) = 0.01$. Also, by an argument similar to the one leading to Eq. (9.2.15), the probability of an error of type II will be $\beta(\delta^*) = \Phi(2.326 - n^{1/2})$, where Φ denotes the c.d.f. of the standard normal distribution. For $n = 1, 25$, and 100, the values of $\beta(\delta^*)$ and k are as follows:

n	$\alpha(\delta^*)$	$\beta(\delta^*)$	k
1	0.01	0.91	6.21
25	0.01	0.0038	0.42
100	0.01	8×10^{-15}	2.5×10^{-12}

It can be seen from this tabulation that when $n = 1$, the null hypothesis H_0 will be rejected only if the likelihood ratio $f_1(\boldsymbol{x})/f_0(\boldsymbol{x})$ exceeds the value $k = 6.21$. In other words, H_0 will not be rejected unless the observed values $x_1, \ldots, x_n$ in the sample are at least 6.21 times as likely under H_1 as they are under H_0. In this case, the procedure δ^* therefore satisfies the experimenter's desire to use a test that is cautious about rejecting H_0.

If $n = 100$, however, the procedure δ^* will reject H_0 whenever the likelihood ratio exceeds the value $k = 2.5 \times 10^{-12}$. Therefore, H_0 will be rejected for certain observed values $x_1, \ldots, x_n$ that are actually millions of times more likely under H_0 as they are under H_1. The reason for this result is that the value of $\beta(\delta^*)$ that can be achieved when $n = 100$, which is 8×10^{-15}, is extremely small relative to the specified value $\alpha_0 = 0.01$. Hence, the procedure δ^* actually turns out to be much more cautious about an error of type II than it is about an error of type I. We can see from this discussion that a value of α_0 that is an appropriate choice for a small value of n might be unnecessarily large for a large value of n. Hence, it would be sensible to let the level of significance α_0 decrease as the sample size increases.

Suppose now that the experimenter regards an error of type I to be much more serious than an error of type II, and she therefore desires to use a test procedure for which the value of the linear combination $100\alpha(\delta) + \beta(\delta)$ will be a minimum. Then it follows from Theorem 9.2.1 that she should reject H_0 if and only if the likelihood ratio exceeds the value $k = 100$, regardless of the sample size n. In other words, the procedure that minimizes the value of $100\alpha(\delta) + \beta(\delta)$ will not reject H_0 unless the observed values $x_1, \ldots, x_n$ are at least 100 times as likely under H_1 as they are under H_0.

From this discussion, it seems more reasonable for the experimenter to take the values of both $\alpha(\delta)$ and $\beta(\delta)$ into account when choosing a test procedure, rather than to fix a value of $\alpha(\delta)$ and minimize $\beta(\delta)$. For example, one could minimize the value of a linear combination of the form $a\alpha(\delta) + b\beta(\delta)$. In Sec. 9.8, we saw how the Bayesian point of view also leads to the conclusion that one should try to minimize a linear combination of this form. Lehmann (1958) suggested choosing a number k and requiring that $\beta(\delta) = k\alpha(\delta)$. Both the Bayesian method and Lehmann's method have the advantage of forcing the probabilities of both type I and type II errors to decrease as one obtains more data. Similar problems with fixing the significance level of a test arise when hypotheses are composite, as we illustrate later in this section.

Statistically Significant Results

When the observed data lead to rejecting a null hypothesis H_0 at level α_0, it is often said that one has obtained a result that is *statistically significant* at level α_0. When this occurs, it does not mean that the experimenter should behave as if H_0 is false. Similarly, if the data do not lead to rejecting H_0, the result is not statistically significant at level α_0, but the experimenter should not necessarily become convinced that H_0 is true. Indeed, qualifying "significant" with the term "statistically" is a warning that a statistically significant result might be different than a practically significant result. Consider, once again, Example 9.5.10 on page 582, in which the hypotheses to be tested are

$$H_0: \quad \mu = 5.2,$$
$$H_1: \quad \mu \neq 5.2.$$

It is extremely important for the experimenter to distinguish a statistically significant result from any claim that the parameter μ is significantly different from the hypothesized value 5.2. Even if the data suggest that μ is not equal to 5.2, this does not necessarily provide any evidence that the actual value of μ is *significantly* different from 5.2. For a given set of data, the tail area corresponding to the observed value of the test statistic U might be very small, and yet the data might suggest that the actual value of μ is so close to 5.2 that, for practical purposes, the experimenter would not regard μ as being significantly different from 5.2.

The situation just described can arise when the statistic U is based on a very large random sample. Suppose, for instance, that in Example 9.5.10 the lengths of 20,000 fibers in a random sample are measured, rather than the lengths of only 15 fibers. For a given level of significance, say, $\alpha_0 = 0.05$, let $\pi(\mu, \sigma^2|\delta)$ denote the power function of the t test based on these 20,000 observations. Then $\pi(5.2, \sigma^2|\delta) = 0.05$ for every value of $\sigma^2 > 0$. However, because of the very large number of observations on which the test is based, the power $\pi(\mu, \sigma^2|\delta)$ will be very close to 1 for each value of μ that differs only slightly from 5.2 and for a moderate value of σ^2. In other words, even if the value of μ differs only slightly from 5.2, the probability is close to 1 that one

would obtain a statistically significant result. For example, with $n = 20{,}000$, the power of the level 0.05 test when $|\mu - 5.2| = 0.03\sigma$ is 0.99.

As explained in Sec. 9.4, it is inconceivable that the mean length μ of all the fibers in the entire population will be exactly 5.2. However, μ may be very close to 5.2, and when it is, the experimenter will not want to reject the null hypothesis H_0. Nevertheless, it is very likely that the t test based on the sample of 20,000 fibers will lead to a statistically significant result. Therefore, when an experimenter analyzes a powerful test based on a very large sample, he must exercise caution in interpreting the actual significance of a "statistically significant" result. He knows in advance that there is a high probability of rejecting H_0 even when the true value of μ differs only slightly from the value 5.2 specified under H_0.

One way to handle this situation, as discussed earlier in this section, is to recognize that a level of significance much smaller than the traditional value of 0.05 or 0.01 is appropriate for a problem with a large sample size. Another way is to replace the single value of μ in the null hypothesis by an interval, as we did on pages 571 and 610. A third way is to regard the statistical problem as one of estimation rather than one of testing hypotheses.

When a large random sample is available, the sample mean and the sample variance will be excellent estimators of the parameters μ and σ^2. Before the experimenter chooses any decision involving the unknown values of μ and σ^2, she should calculate and consider the values of these estimators as well as the value of the statistic U.

Summary

When we reject a null hypothesis, we say that we have obtained a statistically significant result. The power function of a level α_0 test becomes very large, even for parameter values close to the null hypothesis, as the size of the sample increases. For the case of simple hypotheses, the probability of type II error can become very small while the probability of type I error stays as large as α_0. One way to avoid this is to let the level of significance decrease as the sample size increases. If one rejects a null hypothesis at a particular level of significance α_0, one must be careful to check whether the data actually suggest any deviation of practical importance from the null hypothesis.

Exercises

1. Suppose that a single observation X is taken from the normal distribution with unknown mean μ and known variance is 1. Suppose that it is known that the value of μ must be -5, 0, or 5, and it is desired to test the following hypotheses at the level of significance 0.05:

$$H_0:\ \mu = 0,$$
$$H_1:\ \mu = -5 \text{ or } \mu = 5.$$

Suppose also that the test procedure to be used specifies rejecting H_0 when $|X| > c$, where the constant c is chosen so that $\Pr(|X| > c|\mu = 0) = 0.05$.

a. Find the value of c, and show that if $X = 2$, then H_0 will be rejected.

b. Show that if $X = 2$, then the value of the likelihood function at $\mu = 0$ is 12.2 times as large as its value at $\mu = 5$ and is 5.9×10^9 times as large as its value at $\mu = -5$.

2. Suppose that a random sample of 10,000 observations is taken from the normal distribution with unknown mean μ and known variance is 1, and it is desired to test the following hypotheses at the level of significance 0.05:

$$H_0:\ \mu = 0,$$
$$H_1:\ \mu \neq 0.$$

Suppose also that the test procedure specifies rejecting H_0 when $|\overline{X}_n| \geq c$, where the constant c is chosen so that $\Pr(|\overline{X}_n| \geq c|\mu = 0) = 0.05$. Find the probability that the

test will reject H_0 if **(a)** the actual value of μ is 0.01, and **(b)** the actual value of μ is 0.02.

3. Consider again the conditions of Exercise 2, but suppose now that it is desired to test the following hypotheses:

$$H_0: \quad \mu \le 0,$$
$$H_1: \quad \mu > 0.$$

Suppose also that in the random sample of 10,000 observations, the sample mean $\overline{X}_n$ is 0.03. At what level of significance is this result just significant?

4. Suppose that $X_1, \dots, X_n$ comprise a random sample from the normal distribution with unknown mean θ and known variance 1. Suppose that it is desired to test the same hypotheses as in Exercise 3. This time, however, the test procedure δ will be chosen so as to minimize $19\pi(0|\delta) + 1 - \pi(0.5|\delta)$.

a. Find the value c_n so that the test procedure δ rejects H_0 if $\overline{X}_n \ge c_n$ for each value $n = 1$, $n = 100$, and $n = 10{,}000$.

b. For each value of n in part (a), find the size of the test procedure δ.

5. Suppose that $X_1, \dots, X_n$ comprise a random sample from the normal distribution with unknown mean θ and variance 1. Suppose that it is desired to test the same hypotheses as in Exercise 3. This time, however, the test procedure δ will be chosen so that $19\pi(0|\delta) = 1 - \pi(0.5|\delta)$.

a. Find the value c_n so that the test procedure δ rejects H_0 if $\overline{X}_n \ge c_n$ for each value $n = 1$, $n = 100$, and $n = 10{,}000$.

b. For each value of n in part (a), find the size of the test procedure δ.

9.10 Supplementary Exercises

1. I will flip a coin three times and let X stand for the number of times that the coin comes up heads. Let θ stand for the probability that the coin comes up heads on a single flip, and assume that the flips are independent given θ. I wish to test the null hypothesis $H_0: \theta = 1/2$ against the alternative hypothesis $H_1: \theta = 3/4$. Find the test δ that minimizes $\alpha(\delta) + \beta(\delta)$, the sum of the type I and type II error probabilities, and find the two error probabilities for the test.

2. Suppose that a sequence of Bernoulli trials is to be carried out with an unknown probability θ of success on each trial, and the following hypotheses are to be tested:

$$H_0: \quad \theta = 0.1,$$
$$H_1: \quad \theta = 0.2.$$

Let X denote the number of trials required to obtain a success, and suppose that H_0 is to be rejected if $X \le 5$. Determine the probabilities of errors of type I and type II.

3. Consider again the conditions of Exercise 2. Suppose that the losses from errors of type I and type II are equal, and the prior probabilities that H_0 and H_1 are true are equal. Determine the Bayes test procedure based on the observation X.

4. Suppose that a single observation X is to be drawn from the following p.d.f.:

$$f(x|\theta) = \begin{cases} 2(1-\theta)x + \theta & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

where the value of θ is unknown $(0 \le \theta \le 2)$. Suppose also that the following hypotheses are to be tested:

$$H_0: \quad \theta = 2,$$
$$H_1: \quad \theta = 0.$$

Determine the test procedure δ for which $\alpha(\delta) + 2\beta(\delta)$ is a minimum, and calculate this minimum value.

5. Consider again the conditions of Exercise 4, and suppose that $\alpha(\delta)$ is required to be a given value α_0 $(0 < \alpha_0 < 1)$. Determine the test procedure δ for which $\beta(\delta)$ will be a minimum, and calculate this minimum value.

6. Consider again the conditions of Exercise 4, but suppose now that the following hypotheses are to be tested:

$$H_0: \quad \theta \ge 1,$$
$$H_1: \quad \theta < 1.$$

a. Determine the power function of the test δ that specifies rejecting H_0 if $X \ge 0.9$.

b. What is the size of the test δ?

7. Consider again the conditions of Exercise 4. Show that the p.d.f. $f(x|\theta)$ has a monotone likelihood ratio in the statistic $r(X) = -X$, and determine a UMP test of the following hypotheses at the level of significance $\alpha_0 = 0.05$:

$$H_0: \quad \theta \le \tfrac{1}{2},$$
$$H_1: \quad \theta > \tfrac{1}{2}.$$

8. Suppose that a box contains a large number of chips of three different colors, red, brown, and blue, and it is desired to test the null hypothesis H_0 that chips of the three colors are present in equal proportions against the alternative hypothesis H_1 that they are not present in equal proportions. Suppose that three chips are to be drawn at

random from the box, and H_0 is to be rejected if and only if at least two of the chips have the same color.

a. Determine the size of the test.

b. Determine the power of the test if 1/7 of the chips are red, 2/7 are brown, and 4/7 are blue.

9. Suppose that a single observation X is to be drawn from an unknown distribution P, and that the following simple hypotheses are to be tested:

H_0: P is the uniform distribution on the interval [0, 1],

H_1: P is the standard normal distribution.

Determine the most powerful test of size 0.01, and calculate the power of the test when H_1 is true.

10. Suppose that the 12 observations $X_1, \ldots, X_{12}$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Describe how to carry out a t test of the following hypotheses at the level of significance $\alpha_0 = 0.005$:

$$H_0: \quad \mu \geq 3,$$
$$H_1: \quad \mu < 3.$$

11. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean θ and known variance 1, and it is desired to test the following hypotheses:

$$H_0: \quad \theta \leq 0,$$
$$H_1: \quad \theta > 0.$$

Suppose also that it is decided to use a UMP test for which the power is 0.95 when $\theta = 1$. Determine the size of this test if $n = 16$.

12. Suppose that eight observations $X_1, \ldots, X_8$ are drawn at random from a distribution with the following p.d.f.:

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose also that the value of θ is unknown ($\theta > 0$), and it is desired to test the following hypotheses:

$$H_0: \quad \theta \leq 1,$$
$$H_1: \quad \theta > 1.$$

Show that the UMP test at the level of significance $\alpha_0 = 0.05$ specifies rejecting H_0 if $\sum_{i=1}^{8} \log X_i \geq -3.981$.

13. Suppose that $X_1, \ldots, X_n$ form a random sample from the χ^2 distribution with unknown degrees of freedom θ ($\theta = 1, 2, \ldots$), and it is desired to test the following hypotheses at a given level of significance α_0 ($0 < \alpha_0 < 1$):

$$H_0: \quad \theta \leq 8,$$
$$H_1: \quad \theta \geq 9.$$

Show that there exists a UMP test, and the test specifies rejecting H_0 if $\sum_{i=1}^{n} \log X_i \geq k$ for some appropriate constant k.

14. Suppose that $X_1, \ldots, X_{10}$ form a random sample from a normal distribution for which both the mean and the variance are unknown. Construct a statistic that does not depend on any unknown parameters and has the F distribution with three and five degrees of freedom.

15. Suppose that $X_1, \ldots, X_m$ form a random sample from the normal distribution with unknown mean μ_1 and unknown variance σ_1^2, and that $Y_1, \ldots, Y_n$ form an independent random sample from the normal distribution with unknown mean μ_2 and unknown variance σ_2^2. Suppose also that it is desired to test the following hypotheses with the usual F test at the level of significance $\alpha_0 = 0.05$:

$$H_0: \quad \sigma_1^2 \leq \sigma_2^2,$$
$$H_1: \quad \sigma_1^2 > \sigma_2^2.$$

Assuming that $m = 16$ and $n = 21$, show that the power of the test when $\sigma_1^2 = 2\sigma_2^2$ is given by $\Pr(V^* \geq 1.1)$, where V^* is a random variable having the F distribution with 15 and 20 degrees of freedom.

16. Suppose that the nine observations $X_1, \ldots, X_9$ form a random sample from the normal distribution with unknown mean μ_1 and unknown variance σ^2, and the nine observations $Y_1, \ldots, Y_9$ form an independent random sample from the normal distribution with unknown mean μ_2 and the same unknown variance σ^2. Let S_X^2 and S_Y^2 be as defined in Eq. (9.6.2) (with $m = n = 9$), and let

$$T = \max\left\{\frac{S_X^2}{S_Y^2}, \frac{S_Y^2}{S_X^2}\right\}.$$

Determine the value of the constant c such that $\Pr(T > c) = 0.05$.

17. An unethical experimenter desires to test the following hypotheses:

$$H_0: \quad \theta = \theta_0,$$
$$H_1: \quad \theta \neq \theta_0.$$

She draws a random sample $X_1, \ldots, X_n$ from a distribution with the p.d.f. $f(x|\theta)$, and carries out a test of size α. If this test does not reject H_0, she discards the sample, draws a new independent random sample of n observations, and repeats the test based on the new sample. She continues drawing new independent samples in this way until she obtains a sample for which H_0 is rejected.

a. What is the overall size of this testing procedure?

b. If H_0 is true, what is the expected number of samples that the experimenter will have to draw until she rejects H_0?

18. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown precision τ, and the following hypotheses are to

be tested:

$$H_0:\quad \mu \le 3,$$
$$H_1:\quad \mu > 3.$$

Suppose that the prior joint distribution of μ and τ is the normal-gamma distribution, as described in Theorem 8.6.1, with $\mu_0 = 3$, $\lambda_0 = 1$, $\alpha_0 = 1$, and $\beta_0 = 1$. Suppose finally that $n = 17$, and it is found from the observed values in the sample that $\overline{X}_n = 3.2$ and $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 17$. Determine both the prior probability and the posterior probability that H_0 is true.

19. Consider a problem of testing hypotheses in which the following hypotheses about an arbitrary parameter θ are to be tested:

$$H_0:\quad \theta \in \Omega_0,$$
$$H_1:\quad \theta \in \Omega_1.$$

Suppose that δ is a test procedure of size α $(0 < \alpha < 1)$ based on some vector of observations $\boldsymbol{X}$, and let $\pi(\theta|\delta)$ denote the power function of δ. Show that if δ is unbiased, then $\pi(\theta|\delta) \ge \alpha$ at every point $\theta \in \Omega_1$.

20. Consider again the conditions of Exercise 19. Suppose now that we have a two-dimensional vector $\boldsymbol{\theta} = (\theta_1, \theta_2)$, where θ_1 and θ_2 are real-valued parameters. Suppose also that A is a particular circle in the $\theta_1\theta_2$-plane, and that the hypotheses to be tested are as follows:

$$H_0:\quad \theta \in A,$$
$$H_1:\quad \theta \notin A.$$

Show that if the test procedure δ is unbiased and of size α, and if its power function $\pi(\theta|\delta)$ is a continuous function of θ, then it must be true that $\pi(\theta|\delta) = \alpha$ at each point θ on the boundary of the circle A.

21. Consider again the conditions of Exercise 19. Suppose now that θ is a real-valued parameter, and the following hypotheses are to be tested:

$$H_0:\quad \theta = \theta_0,$$
$$H_1:\quad \theta \ne \theta_0.$$

Assume that θ_0 is an interior point of the parameter space Ω. Show that if the test procedure δ is unbiased and if its power function $\pi(\theta|\delta)$ is a differentiable function of θ, then $\pi'(\theta_0|\delta) = 0$, where $\pi'(\theta_0|\delta)$ denotes the derivative of $\pi(\theta|\delta)$ evaluated at the point $\theta = \theta_0$.

22. Suppose that the differential brightness θ of a certain star has an unknown value, and it is desired to test the following simple hypotheses:

$$H_0:\quad \theta = 0,$$
$$H_1:\quad \theta = 10.$$

The statistician knows that when he goes to the observatory at midnight to measure θ, there is probability 1/2 that the meteorological conditions will be good, and he will be able to obtain a measurement X having the normal distribution with mean θ and variance 1. He also knows that there is probability 1/2 that the meteorological conditions will be poor, and he will obtain a measurement Y having the normal distribution with mean θ and variance 100. The statistician also learns whether the meteorological conditions were good or poor.

- **a.** Construct the most powerful test that has conditional size $\alpha = 0.05$, given good meteorological conditions, and one that has conditional size $\alpha = 0.05$, given poor meteorological conditions.
- **b.** Construct the most powerful test that has conditional size $\alpha = 2.0 \times 10^{-7}$, given good meteorological conditions, and one that has conditional size $\alpha = 0.0999998$, given poor meteorological conditions. (You will need a computer program to do this.)
- **c.** Show that the overall size of both the test found in part (a) and the test found in part (b) is 0.05, and determine the power of each of these two tests.

23. Consider again the situation described in Exercise 22. This time, assume that there is a loss function of the form (9.8.6). Also, assume that the prior probability of $\theta = 0$ is ξ_0 and the prior probability of $\theta = 10$ is ξ_1.

- **a.** Find the formula for the Bayes test for general loss function of the form (9.8.6).
- **b.** Prove that the test in part (a) of Exercise 22 is not a special case of the Bayes test found in part (a) of the present exercise.
- **c.** Prove that the test in part (b) of Exercise 22 is (up to rounding error) a special case of the Bayes test found in part (a) of the present exercise.

24. Let $X_1, \dots, X_n$ be i.i.d. with the Poisson distribution having mean θ. Let $Y = \sum_{i=1}^{n} X_i$.

- **a.** Suppose that we wish to test the hypotheses $H_0 : \theta \ge 1$ versus $H_1 : \theta < 1$. Show that the test "reject H_0 if $Y = 0$" is uniformly most powerful level α_0 for some number α_0. Also find α_0.
- **b.** Find the power function of the test from part (a).

25. Consider a family of distributions with parameter θ and monotone likelihood ratio in a statistic T. We learned how to find a uniformly most powerful level α_0 test δ_c of the null hypothesis $H_{0,c} : \theta \le c$ versus $H_{1,c} : \theta > c$ for every c. We also know that these tests are equivalent to a coefficient $1 - \alpha_0$ confidence interval, where the confidence interval contains c if and only if δ_c does not reject $H_{0,c}$. The confidence interval is called *uniformly most accurate coefficient* $1 - \alpha_0$. Based on the equivalence of the tests and the confidence interval, figure out what the definition of "uniformly most accurate coefficient $1 - \alpha_0$" must be. Write the definition in terms of the conditional probability that the interval covers θ_1 given that $\theta = \theta_2$ for various pairs of values θ_1 and θ_2.

Chapter 10

Categorical Data and Nonparametric Methods

10.1 Tests of Goodness-of-Fit

In some problems, we have one specific distribution in mind for the data we will observe. If that one distribution is not appropriate, we do not necessarily have a parametric family of alternative distributions in mind. In these cases, and others, we can still test the null hypothesis that the data come from the one specific distribution against the alternative hypothesis that the data do not come from that distribution.

Description of Nonparametric Problems

Example 10.1.1 Failure Times of Ball Bearings. In Example 5.6.9, we observed the failure times of 23 ball bearings, and we modeled the logarithms of these failure times as normal random variables. Suppose that we are not so confident that the normal distribution is a good model for the logarithms of the failure times. Is there a way to test the null hypothesis that a normal distribution is a good model against the alternative that no normal distribution is a good model? Is there a way to estimate features of the distribution of failure times (such as the median, variance, etc.) if we are unwilling to model the data as normal random variables? ◄

In each of the problems of estimation and testing hypotheses that we considered in Chapters 7, 8, and 9, we have assumed that the observations that are available to the statistician come from distributions for which the exact form is known, even though the values of some parameters are unknown. For example, it might be assumed that the observations form a random sample from a Poisson distribution for which the mean is unknown, or it might be assumed that the observations come from two normal distributions for which the means and variances are unknown. In other words, we have assumed that the observations come from a certain *parametric family* of distributions, and a statistical inference must be made about the values of the parameters defining that family.

In many of the problems to be discussed in this chapter, we shall not assume that the available observations come from a particular parametric family of distributions. Rather, we shall study inferences that can be made about the distribution from which the observations come, without making special assumptions about the form of that distribution. As one example, we might simply assume that the observations form

a random sample from a continuous distribution, without specifying the form of this distribution any further, and we might then investigate the possibility that this distribution is a normal distribution. As a second example, we might be interested in making an inference about the value of the median of the distribution from which the sample was drawn, and we might assume only that this distribution is continuous. As a third example, we might be interested in investigating the possibility that two independent random samples actually come from the same distribution, and we might assume only that both distributions from which the samples are taken are continuous.

Problems in which the possible distributions of the observations are not restricted to a specific parametric family are called *nonparametric problems*, and the statistical methods that are applicable in such problems are called *nonparametric methods*.

Categorical Data

Example 10.1.2 **Blood Types.** In Example 5.9.3, we learned about a study of blood types among a sample of 6004 white Californians. Suppose that the actual counts of people with the four blood types are given in Table 10.1. We might be interested in whether or not these data are consistent with a theory that predicts a particular set of probabilities for the blood types. Table 10.2 gives theoretical probabilities for the four blood types. How can we go about testing the null hypothesis that the theoretical probabilities in Table 10.2 are the probabilities with which the data in Table 10.1 were sampled? ◄

In this section and the next four sections, we shall consider statistical problems based on data such that each observation can be classified as belonging to one of a finite number of possible categories or types. Observations of this type are called *categorical data*. Since there are only a finite number of possible categories in these problems, and since we are interested in making inferences about the probabilities of these categories, these problems actually involve just a finite number of parameters. However, as we shall see, methods based on categorical data can be usefully applied in both parametric and nonparametric problems.

Table 10.1 Counts of blood types for white Californians

A	B	AB	O
2162	738	228	2876

Table 10.2 Theoretical probabilities of blood types for white Californians

A	B	AB	O
1/3	1/8	1/24	1/2

The χ^2 Test

Suppose that a large population consists of items of k different types, and let p_i denote the probability that an item selected at random will be of type i ($i = 1, \dots, k$). Example 10.1.2 is of this type with $k = 4$. Of course, $p_i \ge 0$ for $i = 1, \dots, k$ and $\sum_{i=1}^k p_i = 1$. Let $p_1^0, \dots, p_k^0$ be specific numbers such that $p_i^0 > 0$ for $i = 1, \dots, k$ and $\sum_{i=1}^k p_i^0 = 1$, and suppose that the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad p_i = p_i^0 \quad \text{for } i = 1, \dots, k, \\ H_1&: \quad p_i \ne p_i^0 \quad \text{for at least one value of } i. \end{aligned} \tag{10.1.1}$$

We shall assume that a random sample of size n is to be taken from the given population. That is, n independent observations are to be taken, and there is probability p_i that each observation will be of type i ($i = 1, \dots, k$). On the basis of these n observations, the hypotheses (10.1.1) are to be tested.

For $i = 1, \dots, k$, we shall let N_i denote the number of observations in the random sample that are of type i. Thus, $N_1, \dots, N_k$ are nonnegative integers such that $\sum_{i=1}^k N_i = n$. Indeed, $(N_1, \dots, N_n)$ has the multinomial distribution (see Sec. 5.9) with parameters n and $\boldsymbol{p} = (p_1, \dots, p_k)$. When the null hypothesis H_0 is true, the expected number of observations of type i is np_i^0 ($i = 1, \dots, k$). The difference between the actual number of observations N_i and the expected number np_i^0 will tend to be smaller when H_0 is true than when H_0 is not true. It seems reasonable, therefore, to base a test of the hypotheses (10.1.1) on values of the differences $N_i - np_i^0$ for $i = 1, \dots, k$ and reject H_0 when the magnitudes of these differences are relatively large.

In 1900, Karl Pearson proved the following result, whose proof will not be given here.

Theorem 10.1.1 χ^2 Statistic. The following statistic

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0} \tag{10.1.2}$$

has the property that if H_0 is true and the sample size $n \to \infty$, then Q converges in distribution to the χ^2 distribution with $k - 1$ degrees of freedom. (See Definition 6.3.1.) ■

Theorem 10.1.1 says that if H_0 is true and the sample size n is large, the distribution of Q will be approximately the χ^2 distribution with $k - 1$ degrees of freedom. The discussion that we have presented indicates that H_0 should be rejected when $Q \ge c$, where c is an appropriate constant. If it is desired to carry out the test at the level of significance α_0, then c should be chosen to be the $1 - \alpha_0$ quantile of the χ^2 distribution with $k - 1$ degrees of freedom. This test is called the χ^2 *test of goodness-of-fit*.

Note: General form of χ^2 test statistic. The form of the statistic Q in (10.1.2) is common to all χ^2 tests including those that will be introduced later in this chapter. The form is a sum of terms, each of which is the square of the difference between an observed count and an expected count divided by the expected count: $\sum(\text{observed}-\text{expected})^2/\text{expected}$. The expected counts are computed under the assumption that the null hypothesis is true.

Whenever the value of each expected count, np_i^0 ($i = 1, \dots, k$), is not too small, the χ^2 distribution will be a good approximation to the actual distribution of Q.

Specifically, the approximation will be very good if $np_i^0 \ge 5$ for $i = 1, \ldots, k$, and the approximation should still be satisfactory if $np_i^0 \ge 1.5$ for $i = 1, \ldots, k$.

We shall now illustrate the use of the χ^2 test of goodness-of-fit by some examples.

Example 10.1.3 Blood Types. In Example 10.1.2, we have specified a hypothetical vector of probabilities $(p_1^0, \ldots, p_4^0)$ for the four blood types in Table 10.2. We can use the data in Table 10.1 to test the null hypothesis H_0 that the probabilities $(p_1, \ldots, p_4)$ of the four blood types equal $(p_1^0, \ldots, p_4^0)$. The four expected counts under H_0 are

$$np_1^0 = 6004 \times \frac{1}{3} = 2001.3, \qquad np_2^0 = 6004 \times \frac{1}{8} = 750.5,$$

$$np_3^0 = 6004 \times \frac{1}{24} = 250.2, \quad \text{and} \quad np_4^0 = 6004 \times \frac{1}{2} = 3002.0.$$

The χ^2 test statistic is then

$$Q = \frac{(2162 - 2001.3)^2}{2001.3} + \frac{738 - 750.5}{750.5} + \frac{(228 - 250.2)^2}{250.2} + \frac{(2876 - 3002.0)^2}{3002.0} = 20.37.$$

To test H_0 at level α_0, we would compare Q to the $1 - \alpha_0$ quantile of the χ^2 distribution with three degrees of freedom. Alternatively, we can compute the p-value, which would be the smallest α_0 at which we could reject H_0. In the case of the χ^2 goodness of fit test, the p-value equals $1 - \mathrm{X}^2_{k-1}(Q)$, where X^2_{k-1} is the c.d.f. of the χ^2 distribution with $k - 1$ degrees of freedom. In this example, $k = 4$ and the p-value is 1.42×10^{-4}. ◀

Example 10.1.4 Montana Outlook Poll. The Bureau of Business and Economic Research at the University of Montana conducted a poll of opinions of Montana residents in May 1992. Among other things, respondents were asked whether their personal financial status was worse, the same, or better than one year ago. Table 10.3 displays some results. We might be interested in whether the respondents' answers are uniformly distributed over the three possible responses. That is, we can test the null hypothesis that the probabilities of the three responses are all equal to 1/3. We calculate

$$Q = \frac{(58 - 189/3)^2}{189/3} + \frac{(64 - 189/3)^2}{189/3} + \frac{(67 - 189/3)^2}{189/3} = 0.6667.$$

Since 0.6667 is the 0.283 quantile of the χ^2 distribution with two degrees of freedom, we would only reject the null at levels greater than $1 - 0.283 = 0.717$. ◀

Example 10.1.5 Testing Hypotheses about a Proportion. Suppose that the proportion p of defective items in a large population of manufactured items is unknown and that the following

Table 10.3 Responses to personal financial status question from Montana Outlook Poll

Worse	Same	Better	Total
58	64	67	189

hypotheses are to be tested:

$$\begin{aligned} H_0:&\quad p = 0.1,\\ H_1:&\quad p \neq 0.1. \end{aligned} \tag{10.1.3}$$

Suppose also that in a random sample of 100 items, it is found that 16 are defective. We shall test the hypotheses (10.1.3) by carrying out a χ^2 test of goodness-of-fit.

Since there are only two types of items in this example, namely, defective items and nondefective items, we know that $k = 2$. Furthermore, if we let p_1 denote the unknown proportion of defective items and let p_2 denote the unknown proportion of nondefective items, then the hypotheses (10.1.3) can be rewritten in the following form:

$$\begin{aligned} H_0:&\quad p_1 = 0.1 \text{ and } p_2 = 0.9,\\ H_1:&\quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{10.1.4}$$

For the sample size $n = 100$, the expected number of defective items if H_0 is true is $np_1^0 = 10$, and the expected number of nondefective items is $np_2^0 = 90$. Let N_1 denote the number of defective items in the sample, and let N_2 denote the number of nondefective items in the sample. Then, when H_0 is true, the distribution of the statistic Q defined by Eq. (10.1.2) will be approximately the χ^2 distribution with one degree of freedom.

In this example, $N_1 = 16$ and $N_2 = 84$, and it is found that the value of Q is 4. It can now be determined, either from interpolation in a table of the χ^2 distribution with one degree of freedom or from statistical software, that the tail area (p-value) corresponding to the value $Q = 4$ is approximately 0.0455. Hence, the null hypothesis H_0 would be rejected at levels of significance greater than 0.0455, but not at smaller levels. For hypotheses about a single proportion, we developed tests in Sec. 9.1. (See Exercise 11 in Sec. 9.1, for example.) You can compare the test from Sec. 9.1 to the test in this example in Exercise 1 at the end of this section. ◀

Testing Hypotheses about a Continuous Distribution

Consider a random variable X that takes values in the interval $0 < X < 1$ but has an unknown p.d.f. over this interval. Suppose that a random sample of 100 observations is taken from this unknown distribution, and it is desired to test the null hypothesis that the distribution is the uniform distribution on the interval [0, 1] against the alternative hypothesis that the distribution is not uniform. This problem is a nonparametric problem, since the distribution of X might be any continuous distribution on the interval [0, 1]. However, as we shall now show, the χ^2 test of goodness-of-fit can be applied to this problem.

Suppose that we divide the interval [0, 1] into 20 subintervals of equal length, namely, the interval [0, 0.05), the interval [0.05, 0.10), and so on. If the actual distribution is a uniform distribution, then the probability that each observation will fall within the ith subinterval is 1/20, for $i = 1, \ldots, 20$. Since the sample size in this example is $n = 100$, it follows that the expected number of observations in each subinterval is 5. If N_i denotes the number of observations in the sample that actually fall within the ith subinterval, then the statistic Q defined by Eq. (10.1.2) can be rewritten simply as follows:

$$Q = \frac{1}{5}\sum_{i=1}^{20}(N_i - 5)^2. \tag{10.1.5}$$

If the null hypothesis is true, and the distribution from which the observations were taken is indeed a uniform distribution, then Q will have approximately the χ^2 distribution with 19 degrees of freedom.

The method that has been presented in this example obviously can be applied to every continuous distribution. To test whether a random sample of observations comes from a particular distribution, the following procedure can be adopted:

i. Partition the entire real line, or any particular interval that has probability 1, into a finite number k of disjoint subintervals. Generally, k is chosen so that the expected number of observations in each subinterval is at least 5 if H_0 is true.

ii. Determine the probability p_i^0 that the particular hypothesized distribution would assign to the ith subinterval, and calculate the expected number np_i^0 of observations in the ith subinterval ($i = 1, \dots, k$).

iii. Count the number N_i of observations in the sample that fall within the ith subinterval ($i = 1, \dots, k$).

iv. Calculate the value of Q as defined by Eq. (10.1.2). If the hypothesized distribution is correct, Q will have approximately the χ^2 distribution with $k - 1$ degrees of freedom.

Example 10.1.6

Failure Times of Ball Bearings. Return to Example 10.1.1. Suppose that we wish to use the χ^2 test to test the null hypothesis that the logarithms of the lifetimes are an i.i.d. sample from the normal distribution with mean $\log(50) = 3.912$ and variance 0.25. In order to have the expected count in each interval be at least 5, we can use at most $k = 4$ intervals. We shall make these intervals each have probability 0.25 under the null hypothesis. That is, we shall divide the intervals at the 0.25, 0.5, and 0.75 quantiles of the hypothesized normal distribution. These quantiles are

$$
\begin{aligned}
3.912 + 0.5\Phi^{-1}(0.25) &= 3.192 + 0.5 \times (-0.674) = 3.575,\\
3.912 + 0.5\Phi^{-1}(0.5) &= 3.192 + 0.5 \times 0 = 3.912,\\
3.912 + 0.5\Phi^{-1}(0.75) &= 3.192 + 0.5 \times 0.674 = 4.249,
\end{aligned}
$$

because the 0.25 and 0.75 quantiles of the standard normal distribution are ± 0.674. The observed logarithms are

2.88	3.36	3.50	3.73	3.74	3.82	3.88	3.95
3.95	3.99	4.02	4.22	4.23	4.23	4.23	4.43
4.53	4.59	4.66	4.66	4.85	4.85	5.16	

The numbers of observations in each of the four intervals are then 3, 4, 8, and 8. We then calculate

$$
\begin{aligned}
Q = {} & \frac{(3 - 23 \times 0.25)^2}{23 \times 0.25} + \frac{(4 - 23 \times 0.25)^2}{23 \times 0.25} + \frac{(8 - 23 \times 0.25)^2}{23 \times 0.25}\\
& + \frac{(8 - 23 \times 0.25)^2}{23 \times 0.25} = 3.609.
\end{aligned}
$$

Our table of the χ^2 distribution with three degrees of freedom indicates that 3.609 is between the 0.6 and 0.7 quantiles, so we would not reject the null hypothesis at levels less 0.3 and reject the null hypothesis at levels greater than 0.4. (Actually, the p-value is 0.307.) ◄

One arbitrary feature of the procedure just described is the way in which the subintervals are chosen. Two statisticians working on the same problem might very well choose the subintervals in two different ways. Generally speaking, it is a good policy to choose the subintervals so that the expected numbers of observations in the individual subintervals are approximately equal, and also to choose as many subintervals as possible without allowing the expected number of observations in any subinterval to become small. This is what we did in Example 10.1.6.

Likelihood Ratio Tests for Proportions

In Examples 10.1.3 and 10.1.4, we used the χ^2 goodness-of-fit test to test hypotheses of the form (10.1.4). Although χ^2 tests are commonly used in such examples, we could actually use parametric tests in these examples. For example, the vector of responses in Table 10.3 can be thought of as the observed value of a multinomial random vector with parameters 189 and $\boldsymbol{p} = (p_1, p_2, p_3)$. (See Sec. 5.9.) The hypotheses in Eq. (10.1.4) are then of the form

$$H_0: \boldsymbol{p} = \boldsymbol{p}^{(0)} \text{ versus } H_1: \ H_0 \text{ is not true.}$$

As such, we can use the method of likelihood ratio tests for testing the hypotheses. Specifically, we shall apply Theorem 9.1.4. The likelihood function from a multinomial vector $\boldsymbol{x} = (N_1, \dots, N_k)$ is

$$f(\boldsymbol{x}|\boldsymbol{p}) = \binom{n}{N_1, \dots, N_k} p_1^{N_1} \cdots p_k^{N_k}. \tag{10.1.6}$$

In order to apply Theorem 9.1.4, the parameter space must be an open set in k-dimensional space. This is not true for the multinomial distribution if we let $\boldsymbol{p}$ be the parameter. The set of probability vectors lies in a $(k-1)$-dimensional subset of k-dimensional space because the coordinates are constrained to add up to 1. However, we can just as effectively treat the vector $\theta = (p_1, \dots, p_{k-1})$ as the parameter because $p_k = 1 - p_1 - \cdots - p_{k-1}$ is a function of θ. As long as we believe that all coordinates of $\boldsymbol{p}$ are strictly between 0 and 1, the set of possible values of the $(k-1)$-dimensional parameter θ is open. The likelihood function (10.1.6) can then be rewritten as

$$g(\boldsymbol{x}|\theta) = \binom{n}{N_1, \dots, N_k} \theta_1^{N_1} \cdots \theta_{k-1}^{N_{k-1}} (1 - \theta_1 - \cdots - \theta_{k-1})^{N_k}. \tag{10.1.7}$$

If H_0 is true, there is only one possible value for (10.1.7), namely,

$$\binom{n}{N_1, \dots, N_k} (p_1^{(0)})^{N_1} \cdots (p_k^{(0)})^{N_k},$$

which is then the numerator of the likelihood ratio statistic $\Lambda(\boldsymbol{x})$ from Definition 9.1.11. The denominator of $\Lambda(\boldsymbol{x})$ is found by maximizing (10.1.7). It is not difficult to show that the M.L.E.'s are $\hat{\theta}_i = N_i/n$ for $i = 1, \dots, k-1$. The large-sample likelihood ratio test statistic is then

$$-2 \log \Lambda(\boldsymbol{x}) = -2 \sum_{i=1}^{k} N_i \log \left(\frac{np_i^{(0)}}{N_i} \right).$$

The large-sample test rejects H_0 at level of significance α_0 if this statistic is greater than the $1 - \alpha_0$ quantile of the χ^2 distribution with $k - 1$ degrees of freedom.

Example 10.1.7

Blood Types. Using the data in Table 10.1, we can test the null hypothesis that the vector of probabilities equals the vector of numbers in Table 10.2. The values of $np_i^{(0)}$

for $i = 1, 2, 3, 4$ were already calculated in Example 10.1.3. The test statistic is

$$-2\left[2162\log\left(\frac{2001.3}{2162}\right)+738\log\left(\frac{750.5}{738}\right)+228\log\left(\frac{250.2}{228}\right)+2876\log\left(\frac{3002.0}{2876}\right)\right]$$
$$= 20.16.$$

The p-value is the probability that a χ^2 random variable with three degrees of freedom is greater than 20.16, namely, 1.57×10^{-4}. This is nearly the same as the p-value from the χ^2 test in Example 10.1.3. ◄

Discussion of the Test Procedure

The χ^2 test of goodness-of-fit is subject to the criticisms of tests of hypotheses that were presented in Sec. 9.9. In particular, the null hypothesis H_0 in the χ^2 test specifies the distribution of the observations exactly, but it is not likely that the actual distribution of the observations will be exactly the same as that of a random sample from this specific distribution. Therefore, if the χ^2 test is based on a very large number of observations, we can be almost certain that the tail area corresponding to the observed value of Q will be very small. For this reason, a very small tail area should not be regarded as strong evidence against the hypothesis H_0 without further analysis. Before a statistician concludes that the hypothesis H_0 is unsatisfactory, he should be certain that there exist *reasonable* alternative distributions for which the observed values provide a much better fit. For example, the statistician might calculate the values of the statistic Q for a few reasonable alternative distributions in order to be certain that, for at least one of these distributions, the tail area corresponding to the calculated value of Q is substantially larger than it is for the distribution specified by H_0.

A particular feature of the χ^2 test of goodness-of-fit is that the procedure is designed to test the null hypothesis H_0 that $p_i = p_i^0$ for $i = 1, \dots, k$ against the general alternative that H_0 is not true. If it is desired to use a test procedure that is especially effective for detecting certain types of deviations of the actual values of $p_1, \dots, p_k$ from the hypothesized values $p_1^0, \dots, p_k^0$, then the statistician should design special tests that have higher power for these types of alternatives and lower power for alternatives of lesser interest. This topic will not be discussed in this book.

Because the random variables $N_1, \dots, N_k$ in Eq. (10.1.2) are discrete, the χ^2 approximation to the distribution of Q can sometimes be improved by introducing a correction for continuity of the type described in Sec. 6.4. However, we shall not use the correction in this book.

Summary

The χ^2 test of goodness-of-fit was introduced as a method for testing the null hypothesis that our data form an i.i.d. sample from a specific distribution against the alternative hypothesis that the data have some other distribution. The test is most natural when the specific distribution is discrete. Suppose that there are k possible values for each observation, and we observe N_i with value i for $i = 1, \dots, k$. Suppose that the null hypothesis says that the probability of the ith possible value is p_i^0 for

$i = 1, \dots, k$. Then we compute

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0},$$

where $n = \sum_{i=1}^{k} N_i$ is the sample size. When the null hypothesis says that the data have a continuous distribution, then one must first create a corresponding discrete distribution. One does this by dividing the real line into finitely many (say, k) intervals, calculating the probability of each interval $p_1^0, \dots, p_k^0$, and then pretending as if all we learned from the data were into which intervals each observation fell. This converts the original data into discrete data with k possible values. For example, the value of N_i used in the formula for Q is the number of observations that fell into the ith interval. All of the χ^2 test statistics in this text have the form $\sum(\text{observed}-\text{expected})^2/\text{expected}$, where "observed" stands for an observed count and "expected" stands for the expected value of the observed count under the assumption that the null hypothesis is true.

Exercises

1. Consider the hypotheses being tested in Example 10.1.5. Use a test procedure of the form outlined in Exercise 11 of Sec. 9.1 and compare the result to the numerical result obtained in Example 10.1.5.

2. Show that if $p_i^0 = 1/k$ for $i = 1, \dots, k$, then the statistic Q defined by Eq. (10.1.2) can be written in the form

$$Q = \left(\frac{k}{n} \sum_{i=1}^{k} N_i^2 \right) - n.$$

3. Investigate the "randomness" of your favorite pseudo-random number generator as follows. Simulate 200 pseudo-random numbers between 0 and 1 and divide the unit interval into $k = 10$ intervals of length 0.1 each. Apply the χ^2 test of the hypothesis that each of the 10 intervals has the same probability of containing a pseudo-random number.

4. According to a simple genetic principle, if both the mother and the father of a child have genotype Aa, then there is probability 1/4 that the child will have genotype AA, probability 1/2 that she will have genotype Aa, and probability 1/4 that she will have genotype aa. In a random sample of 24 children having both parents with genotype Aa, it is found that 10 have genotype AA, 10 have genotype Aa, and four have genotype aa. Investigate whether the simple genetic principle is correct by carrying out a χ^2 test of goodness-of-fit.

5. Suppose that in a sequence of n Bernoulli trials, the probability p of success on each trial is unknown. Suppose also that p_0 is a given number in the interval (0, 1), and it is desired to test the following hypotheses:

$$H_0: \quad p = p_0,$$
$$H_1: \quad p \neq p_0.$$

Let $\overline{X}_n$ denote the proportion of successes in the n trials, and suppose that the given hypotheses are to be tested by using a χ^2 test of goodness-of-fit.

a. Show that the statistic Q defined by Eq. (10.1.2) can be written in the form

$$Q = \frac{n(\overline{X}_n - p_0)^2}{p_0(1 - p_0)}.$$

b. Assuming that H_0 is true, prove that as $n \to \infty$, the c.d.f. of Q converges to the c.d.f. of the χ^2 distribution with one degree of freedom. *Hint:* Show that $Q = Z^2$, where it is known from the central limit theorem that Z is a random variable whose c.d.f. converges to the c.d.f. of the standard normal distribution.

6. It is known that 30 percent of small steel rods produced by a standard process will break when subjected to a load of 3000 pounds. In a random sample of 50 similar rods produced by a new process, it was found that 21 of them broke when subjected to a load of 3000 pounds. Investigate the hypothesis that the breakage rate for the new process is the same as the rate for the old process by carrying out a χ^2 test of goodness-of-fit.

7. In a random sample of 1800 observed values from the interval (0, 1), it was found that 391 values were between 0 and 0.2, 490 values were between 0.2 and 0.5, 580 values were between 0.5 and 0.8, and 339 values were between 0.8 and 1. Test the hypothesis that the random sample was drawn from the uniform distribution on the interval [0, 1] by carrying out a χ^2 test of goodness-of-fit at the level of significance 0.01.

8. Suppose that the distribution of the heights of men who reside in a certain large city is the normal distribution for which the mean is 68 inches and the standard deviation is 1 inch. Suppose also that when the heights of 500 men who reside in a certain neighborhood of the city were measured, the distribution in Table 10.4 was obtained. Test the hypothesis that, with regard to height, these 500 men form a random sample from all the men who reside in the city.

Table 10.4 Data for Exercise 8

Height	Number of men
Less than 66 in.	18
Between 66 and 67.5 in.	177
Between 67.5 and 68.5 in.	198
Between 68.5 and 70 in.	102
Greater than 70 in.	5

9. The 50 values in Table 10.5 are intended to be a random sample from the standard normal distribution.

Table 10.5 Data for Exercise 9

−1.28	−1.22	−0.45	−0.35	0.72
−0.32	−0.80	−1.66	1.39	0.38
−1.38	−1.26	0.49	−0.14	−0.85
2.33	−0.34	−1.96	−0.64	−1.32
−1.14	0.64	3.44	−1.67	0.85
0.41	−0.01	0.67	−1.13	−0.41
−0.49	0.36	−1.24	−0.04	−0.11
1.05	0.04	0.76	0.61	−2.04
0.35	2.82	−0.46	−0.63	−1.61
0.64	0.56	−0.11	0.13	−1.81

a. Carry out a χ^2 test of goodness-of-fit by dividing the real line into five intervals, each of which has probability 0.2 under the standard normal distribution.

b. Carry out a χ^2 test of goodness-of-fit by dividing the real line into 10 intervals, each of which has probability 0.1 under the standard normal distribution.

10.2 Goodness-of-Fit for Composite Hypotheses

We can extend the goodness-of-fit test to deal with the case in which the null hypothesis is that the distribution of our data belongs to a particular parametric family. The alternative hypothesis is that the data have a distribution that is not a member of that parametric family. There are two changes to the test procedure in going from the case of a simple null hypothesis to the case of a composite null hypothesis. First, in the test statistic Q, the probabilities p_i^0 are replaced by estimated probabilities based on the parametric family. Second, the degrees of freedom are reduced by the number of parameters.

Composite Null Hypotheses

Example 10.2.1 Failure Times of Ball Bearings. In Example 10.1.6, we tested the null hypothesis that the logarithms of ball bearing lifetimes have the normal distribution with mean 3.912 and variance 0.25. Suppose that we are not even sure that a normal distribution is a good model for the log-lifetimes. Is there a way for us to test the composite null hypothesis that the distribution of log-lifetimes is a member of the normal family? ◀

We shall consider again a large population that consists of items of k different types and again let p_i denote the probability that an item selected at random will be of type i $(i = 1, \dots, k)$. We shall suppose now, however, that instead of testing the simple null hypothesis that the parameters $p_1, \dots, p_k$ have specific values, we are interested in testing the composite null hypothesis that the values of $p_1, \dots, p_k$ belong to some specified subset of possible values. In particular, we shall consider

problems in which the null hypothesis specifies that the parameters $p_1, \dots, p_k$ can actually be represented as functions of a smaller number of parameters.

Example 10.2.2

Genetics. Consider a gene (such as in Example 1.6.4 on page 23) that has two different alleles. Each individual in a given population must have one of three possible genotypes. If the alleles arrive independently from the two parents, and if every parent has the same probability θ of passing the first allele to each offspring, then the probabilities p_1, p_2, and p_3 of the three different genotypes can be represented in the following form:

$$p_1 = \theta^2, \quad p_2 = 2\theta(1-\theta), \quad p_3 = (1-\theta)^2. \tag{10.2.1}$$

Here, the value of the parameter θ is unknown and can lie anywhere in the interval $0 < \theta < 1$. For each value of θ in this interval, it can be seen that $p_i > 0$ for $i = 1, 2,$ or 3, and $p_1 + p_2 + p_3 = 1$. In this problem, a random sample is taken from the population, and the statistician must use the observed numbers of individuals who have each of the three genotypes to determine whether it is reasonable to believe that there is *some* value of θ in the interval $0 < \theta < 1$ such that p_1, p_2, and p_3 can be represented in the hypothesized form (10.2.1).

If a gene has three different alleles, each individual in the population must have one of six possible genotypes. Once again, if the alleles pass independently from the parents, and if each parent has probabilities θ_1 and θ_2 of passing the first and second alleles, respectively, to an offspring, then the probabilities $p_1, \dots, p_6$ of the different genotypes can be represented in the following form, for *some* values of θ_1 and θ_2 such that $\theta_1 > 0$, $\theta_2 > 0$, and $\theta_1 + \theta_2 < 1$:

$$\begin{aligned} &p_1 = \theta_1^2, \quad p_2 = \theta_2^2, \quad p_3 = (1-\theta_1-\theta_2)^2, \quad p_4 = 2\theta_1\theta_2, \\ &p_5 = 2\theta_1(1-\theta_1-\theta_2), \quad p_6 = 2\theta_2(1-\theta_1-\theta_2). \end{aligned} \tag{10.2.2}$$

Again, for all values of θ_1 and θ_2 satisfying the stated conditions, it can be verified that $p_i > 0$ for $i = 1, \dots, 6$ and $\sum_{i=1}^{6} p_i = 1$. On the basis of the observed numbers $N_1, \dots, N_6$ of individuals having each genotype in a random sample, the statistician must decide whether or not to reject the null hypothesis that the probabilities $p_1, \dots, p_6$ can be represented in the form (10.2.2) for some values of θ_1 and θ_2. ◄

In formal terms, in a problem like those in Example 10.2.2, we are interested in testing the hypothesis that for $i = 1, \dots, k$, each probability p_i can be represented as a particular function $\pi_i(\theta)$ of a vector of parameters $\theta = (\theta_1, \dots, \theta_s)$. It is assumed that $s < k - 1$ and no component of the vector θ can be expressed as a function of the other $s - 1$ components. We shall let Ω denote the s-dimensional parameter space of all possible values of θ. Furthermore, we shall assume that the functions $\pi_1(\theta), \dots, \pi_k(\theta)$ always form a feasible set of values of $p_1, \dots, p_k$ in the sense that for every value of $\theta \in \Omega$, $\pi_i(\theta) > 0$ for $i = 1, \dots, k$ and $\sum_{i=1}^{k} \pi_i(\theta) = 1$.

The hypotheses to be tested can be written in the following form:

$$\begin{aligned} H_0&: \text{ There exists a value of } \theta \in \Omega \text{ such that} \\ &\qquad p_i = \pi_i(\theta) \text{ for } i = 1, \dots, k, \\ H_1&: \text{ The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{10.2.3}$$

The assumption that $s < k - 1$ guarantees that the hypothesis H_0 actually restricts the values of $p_1, \dots, p_k$ to a proper subset of the set of all possible values of these probabilities. In other words, as the vector θ runs through all the values in the set Ω,

the vector $[\pi_1(\theta), \ldots, \pi_k(\theta)]$ runs through only a proper subset of the possible values of $(p_1, \ldots, p_k)$.

The χ^2 Test for Composite Null Hypotheses

In order to carry out a χ^2 test of goodness-of-fit of the hypotheses (10.2.3), the statistic Q defined by Eq. (10.1.2) must be modified because the expected number np_i^0 of observations of type i in a random sample of n observations is no longer completely specified by the null hypothesis H_0. The modification that is used is simply to replace np_i^0 by the M.L.E. of this expected number under the assumption that H_0 is true. In other words, if $\hat{\theta}$ denotes the M.L.E. of the parameter vector θ based on the observed numbers $N_1, \ldots, N_k$, then the statistic Q is defined as follows:

$$Q = \sum_{i=1}^{k} \frac{[N_i - n\pi_i(\hat{\theta})]^2}{n\pi_i(\hat{\theta})}. \tag{10.2.4}$$

Again, it is reasonable to base a test of the hypotheses (10.2.3) on this statistic Q by rejecting H_0 if $Q \geq c$, where c is an appropriate constant. In 1924, R. A. Fisher proved the following result, whose precise statement and proof are not given here. (See Schervish 1995, theorem 7.133.)

Theorem 10.2.1 χ^2 Test for Composite Null. Suppose that the null hypothesis H_0 in (10.2.3) is true and certain regularity conditions are satisfied. Then as the sample size $n \to \infty$, the c.d.f. of Q in (10.2.4) converges to the c.d.f. of the χ^2 distribution with $k - 1 - s$ degrees of freedom. ■

When the sample size n is large and the null hypothesis H_0 is true, the distribution of Q will be approximately a χ^2 distribution. To determine the number of degrees of freedom, we must subtract s from the number $k - 1$ used in Sec. 10.1 because we are now estimating the s parameters $\theta_1, \ldots, \theta_s$ when we compare the observed number N_i with the expected number $n\pi_i(\hat{\theta})$ for $i = 1, \ldots, k$. In order that this result will hold, it is necessary to satisfy the following regularity conditions: First, the M.L.E. $\hat{\theta}$ of the vector θ must occur at a point where the partial derivatives of the likelihood function with respect to each of the parameters $\theta_1, \ldots, \theta_s$ equal 0. Furthermore, these partial derivatives must satisfy certain conditions of the type alluded to in Sec. 8.8 when we discussed the asymptotic properties of M.L.E.'s.

Example 10.2.3 Genetics. As examples of the use of the statistic Q defined by Eq. (10.2.4), consider the two types of genetics problems described in Example 10.2.2. In a problem of the first type, $k = 3$, and it is desired to test the null hypothesis H_0 that the probabilities p_1, p_2, and p_3 can be represented in the form (10.2.1) against the alternative H_1 that H_0 is not true. In this problem, $s = 1$. Therefore, when H_0 is true, the distribution of the statistic Q defined by Eq. (10.2.4) will be approximately the χ^2 distribution with one degree of freedom.

In a problem of the second type, $k = 6$, and it is desired to test the null hypothesis H_0 that the probabilities $p_1, \ldots, p_6$ can be represented in the form (10.2.2) against the alternative H_1 that H_0 is not true. In this problem, $s = 2$. Therefore, when H_0 is true, the distribution of Q will be approximately the χ^2 distribution with three degrees of freedom. ◀

Determining the Maximum Likelihood Estimates

When the null hypothesis H_0 in (10.2.3) is true, the likelihood function $L(\theta)$ for the observed numbers $N_1, \ldots, N_k$ will be

$$L(\theta) = \binom{n}{N_1, \ldots, N_k}[\pi_1(\theta)]^{N_1} \cdots [\pi_k(\theta)]^{N_k}. \tag{10.2.5}$$

Thus,

$$\log L(\theta) = \log \binom{n}{N_1, \ldots, N_k} + \sum_{i=1}^{k} N_i \log \pi_i(\theta). \tag{10.2.6}$$

The M.L.E. $\hat{\theta}$ will be the value of θ for which $\log L(\theta)$ is a maximum. The multinomial coefficient in (10.2.6) does not affect the maximization, and we shall ignore it for the remainder of this section.

Example 10.2.4

Genetics. In the first parts of Examples 10.2.2 and 10.2.3, $k = 3$ and H_0 specifies that the probabilities p_1, p_2, and p_3 can be represented in the form (10.2.1). In this case,

$$\begin{aligned} \log L(\theta) &= N_1 \log(\theta^2) + N_2 \log[2\theta(1-\theta)] + N_3 \log[(1-\theta)^2] \\ &= (2N_1 + N_2) \log\theta + (2N_3 + N_2) \log(1-\theta) + N_2 \log 2. \end{aligned} \tag{10.2.7}$$

It can be found by differentiation that the value of θ for which $\log L(\theta)$ is a maximum is

$$\hat{\theta} = \frac{2N_1 + N_2}{2(N_1 + N_2 + N_3)} = \frac{2N_1 + N_2}{2n}. \tag{10.2.8}$$

The value of the statistic Q defined by Eq. (10.2.4) can now be calculated from the observed numbers N_1, N_2, and N_3. As previously mentioned, when H_0 is true and n is large, the distribution of Q will be approximately the χ^2 distribution with one degree of freedom. Hence, the tail area corresponding to the observed value of Q can be found from that χ^2 distribution. ◄

Testing Whether a Distribution Is Normal

Consider now a problem in which a random sample $X_1, \ldots, X_n$ is taken from some continuous distribution for which the p.d.f. is unknown, and it is desired to test the null hypothesis H_0 that this distribution is a normal distribution against the alternative hypothesis H_1 that the distribution is not normal. To perform a χ^2 test of goodness-of-fit in this problem, divide the real line into k subintervals and count the number N_i of observations in the random sample that fall into the ith subinterval ($i = 1, \ldots, k$).

If H_0 is true, and if μ and σ^2 denote the unknown mean and variance of the normal distribution, then the parameter vector θ is the two-dimensional vector $\theta = (\mu, \sigma^2)$. The probability $\pi_i(\theta)$, or $\pi_i(\mu, \sigma^2)$, that an observation will fall within the ith subinterval, is the probability assigned to that subinterval by the normal distribution with mean μ and variance σ^2. In other words, if the ith subinterval is the interval from a_i to b_i, then

$$\begin{aligned} \pi_i(\mu, \sigma^2) &= \int_{a_i}^{b_i} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx \\ &= \Phi\left(\frac{b_i - \mu}{\sigma}\right) - \Phi\left(\frac{a_i - \mu}{\sigma}\right), \end{aligned} \tag{10.2.9}$$

where $\Phi(\cdot)$ is the standard normal c.d.f., and $\Phi(-\infty) = 0$ and $\Phi(\infty) = 1$.

It is important to note that in order to calculate the value of the statistic Q defined by Eq. (10.2.4), the M.L.E.'s $\hat{\mu}$ and $\hat{\sigma}^2$ must be found by using the numbers $N_1, \dots, N_k$ of observations in the different subintervals. The M.L.E.'s should *not* be found by using the observed values of $X_1, \dots, X_n$ themselves. In other words, $\hat{\mu}$ and $\hat{\sigma}^2$ will be the values of μ and σ^2 that maximize the likelihood function

$$L(\mu, \sigma^2) = [\pi_1(\mu, \sigma^2)]^{N_1} \cdots [\pi_k(\mu, \sigma^2)]^{N_k}. \tag{10.2.10}$$

Because of the complicated nature of the function $\pi_i(\mu, \sigma^2)$, as given by Eq. (10.2.9), a lengthy numerical computation would usually be required to determine the values of μ and σ^2 that maximize $L(\mu, \sigma^2)$. On the other hand, we know that the M.L.E.'s of μ and σ^2 based on the n observed values $X_1, \dots, X_n$ in the original sample are simply the sample mean $\overline{X}_n$ and the sample variance S_n^2/n. Furthermore, if the estimators that maximize the likelihood function $L(\mu, \sigma^2)$ are used to calculate the statistic Q, then we know that when H_0 is true, the distribution of Q will be approximately the χ^2 distribution with $k - 3$ degrees of freedom. On the other hand, if the M.L.E.'s $\overline{X}_n$ and S_n^2/n, which are based on the observed values in the original sample, are used to calculate Q, then this χ^2 approximation to the distribution of Q will not be appropriate. Because of the simple nature of the estimators $\overline{X}_n$ and S_n^2/n, we shall use these estimators to calculate Q, but we shall describe how their use modifies the distribution of Q.

In 1954, H. Chernoff and E. L. Lehmann established the following general result, which we shall not prove here.

Theorem 10.2.2 Let $X_1, \dots, X_n$ be a random sample from a distribution with a p-dimensional parameter θ. Let $\hat{\theta}_n$ denote the M.L.E. as defined in Definition 7.5.2. Partition the real line into $k > p + 1$ disjoint intervals $I_1, \dots, I_k$. Let N_i be the number of observations that fall into I_i for $i = 1, \dots, k$. Let $\pi_i(\theta) = \Pr(X_i \in I_i|\theta)$. Let

$$Q' = \sum_{i=1}^{k} \frac{[N_i - n\pi_i(\hat{\theta}_n)]^2}{n\pi_i(\hat{\theta}_n)}. \tag{10.2.11}$$

Assume the regularity conditions needed for asymptotic normality of the M.L.E. Then, as $n \to \infty$, the c.d.f. of Q' converges to a c.d.f. that lies between the c.d.f. of the χ^2 distribution with $k - p - 1$ degrees of freedom and the c.d.f. of the χ^2 distribution with $k - 1$ degrees of freedom. ■

For the case of testing that the distribution is normal, suppose that we use the M.L.E.'s $\overline{X}_n$ and S_n^2/n and calculate the statistic Q' in Eq. (10.2.11) instead of the statistic Q in Eq. (10.2.4). If the null hypothesis H_0 is true, then as $n \to \infty$, the c.d.f. of Q' converges to a c.d.f. that lies between the c.d.f. of the χ^2 distribution with $k - 3$ degrees of freedom and the c.d.f. of the χ^2 distribution with $k - 1$ degrees of freedom. It follows that if the value of Q' is calculated in this simplified way, then the tail area corresponding to this value of Q' is actually larger than the tail area found from a table of the χ^2 distribution with $k - 3$ degrees of freedom. In fact, the appropriate tail area lies somewhere between the tail area found from a table of the χ^2 distribution with $k - 3$ degrees of freedom and the larger tail area found from a table of the χ^2 distribution with $k - 1$ degrees of freedom. Thus, when the value of Q' is calculated in this simplified way, the corresponding tail area will be bounded by two values that can be obtained from a table of the χ^2 distribution.

Example 10.2.5

Failure Times of Ball Bearings. Return to Example 10.2.1. We are now in a position to try to test the composite null hypothesis that the logarithms of ball bearing lifetimes have some normal distribution. We shall divide the real line into the same subintervals that we used in Example 10.1.6, namely, $(-\infty, 3.575]$, $(3.575, 3.912]$, $(3.912, 4.249]$, and $(4.249, \infty)$. The counts for the four intervals are still 3, 4, 8, and 8. We shall use Theorem 10.2.2, which allows us to use the M.L.E.'s based on the original data. This yields $\hat{\mu} = 4.150$ and $\hat{\sigma}^2 = 0.2722$. The probabilities of the four intervals are

$$\pi_1(\hat{\mu}, \hat{\sigma}^2) = \Phi\left(\frac{3.575 - 4.150}{(0.2722)^{1/2}}\right) = 0.1350,$$

$$\pi_2(\hat{\mu}, \hat{\sigma}^2) = \Phi\left(\frac{3.912 - 4.150}{(0.2722)^{1/2}}\right) - \Phi\left(\frac{3.575 - 4.150}{(0.2722)^{1/2}}\right) = 0.1888,$$

$$\pi_3(\hat{\mu}, \hat{\sigma}^2) = \Phi\left(\frac{4.249 - 4.150}{(0.2722)^{1/2}}\right) - \Phi\left(\frac{3.912 - 4.150}{(0.2722)^{1/2}}\right) = 0.2511,$$

$$\pi_4(\hat{\mu}, \hat{\sigma}^2) = 1 - \Phi\left(\frac{4.249 - 4.150}{(0.2722)^{1/2}}\right) = 0.4251.$$

This makes the value of Q' equal to

$$Q' = \frac{(3 - 23 \times 0.1350)^2}{23 \times 0.1350} + \frac{(4 - 23 \times 0.1888)^2}{23 \times 0.1888} + \frac{(8 - 23 \times 0.2511)^2}{23 \times 0.2511} + \frac{(8 - 23 \times 0.4251)^2}{23 \times 0.4251} = 1.211.$$

The tail area corresponding to 1.211 needs to be computed for χ^2 distributions with $k - 1 = 3$ and $k - 3 = 1$ degrees of freedom. For one degree of freedom, the p-value is 0.2711, and for three degrees of freedom the p-value is 0.7504. So, our actual p-value lies in the interval [0.2711, 0.7504]. Although this interval is wide, it tells not to reject H_0 at level α_0 if $\alpha_0 < 0.2711$. ◀

Note: Testing Composite Hypotheses about an Arbitrary Distribution. Theorem 10.2.2 is very general and applies to both continuous and discrete distributions. Suppose, for example, that a random sample of n observations is taken from a discrete distribution for which the possible values are the nonnegative integers $0, 1, 2, \ldots$. Suppose also that it is desired to test the null hypothesis H_0 that this distribution is a Poisson distribution against the alternative hypothesis H_1 that the distribution is not Poisson. Finally, suppose that the nonnegative integers $0, 1, 2, \ldots$ are divided into k classes such that each observation will lie in one of these classes.

It is known from Exercise 5 of Sec. 7.5 that if H_0 is true, then the sample mean $\overline{X}_n$ is the M.L.E. of the unknown mean θ of the Poisson distribution based on the n observed values in the original sample. Therefore, if the estimator $\hat{\theta} = \overline{X}_n$ is used to calculate the statistic Q' defined by Eq. (10.2.11) rather than the Q in Eq. (10.2.4), then the approximate distribution of Q' when H_0 is true lies between the χ^2 distribution with $k - 2$ degrees of freedom and the χ^2 distribution with $k - 1$ degrees of freedom.

Example 10.2.6

Prussian Army Deaths. In Example 7.3.14, we modeled the numbers of deaths by horsekick in Prussian army units as Poisson random variables. Suppose that we wish to test the null hypothesis that the numbers are a random sample from some Poisson

distribution versus the alternative hypothesis that they are not a Poisson random sample. The numbers of counts reported in Example 7.3.14 are repeated here:

Count	0	1	2	3	≥ 4
Number of Observations	144	91	32	11	2

The likelihood function, assuming that the data form a random sample from a Poisson distribution, is proportional (as a function of θ) to $\exp(-280\theta)\theta^{196}$. The M.L.E. is $\hat{\theta}_n = 196/280 = 0.7$. We can use the $k = 5$ classes above to compute the statistic Q'. The five class probabilities are

Count	0	1	2	3	≥ 4
$\pi_i(\hat{\theta}_n)$	0.4966	0.3476	0.1217	0.0283	0.0058

Then

$$Q' = \frac{(144 - 280 \times 0.4966)^2}{280 \times 0.4966} + \frac{(91 - 280 \times 0.3476)^2}{280 \times 0.3476} + \frac{(32 - 280 \times 0.1217)^2}{280 \times 0.1217} + \frac{(11 - 280 \times 0.0283)^2}{280 \times 0.0283} + \frac{(2 - 208 \times 0.0058)^2}{208 \times 0.0005} = 1.979.$$

The tail areas corresponding to the observed Q' and degrees of freedom four and three are, respectively, 0.7396 and 0.5768. We would not be able to reject H_0 at level α_0 for $\alpha_0 < 0.5768$. ◀

Summary

If we want to test the composite hypothesis that our data have a distribution from a parametric family, we must estimate the parameter θ. We do this by first dividing the real numbers into k disjoint intervals. Then we reduce the data to the counts $N_1, \ldots, N_k$ of how many observations fall into each of the k intervals. We then construct the likelihood function $L(\theta) = \prod_{i=1}^{k} \pi_i(\theta)^{N_i}$, where $\pi_i(\theta)$ is the probability that one observation falls into the ith interval. We estimate θ to be the value $\hat{\theta}$ that maximizes $L(\theta)$. We then compute the test statistic $Q = \sum_{i=1}^{k}[N_i - n\pi_i(\hat{\theta})]^2/[n\pi_i(\hat{\theta})]$, which has the form $\sum(\text{observed}-\text{expected})^2/\text{expected}$. In order to test the null hypothesis at level α_0, we compare Q to the $1 - \alpha_0$ quantile of the χ^2 distribution with $k - 1 - s$ degrees of freedom, where s is the dimension of θ. Alternatively, we can find the usual M.L.E. $\hat{\theta}$ based on the original observations. In this case, we need to compare Q to a number between the $1 - \alpha_0$ quantile of the χ^2 distribution with $k - 1 - s$ degrees of freedom and the $1 - \alpha_0$ quantile of the χ^2 distribution with $k - 1$ degrees of freedom.

Exercises

1. The 41 numbers in Table 10.6 are average sulfur dioxide contents over the years 1969–71 (micrograms per cubic meter) measured in the air in 41 U.S. cities. The data appear on pp. 619–620 of Sokal and Rohlf (1981).

a. Test the null hypothesis that these data arise from a normal distribution.

b. Test the null hypothesis that these data arise from a lognormal distribution.

Table 10.6 Sulfur dioxide in the air of 41 U.S. cities

10	13	12	17	56	36	29
14	10	24	110	28	17	8
30	9	47	35	29	14	56
14	11	46	11	23	65	26
69	61	94	10	18	9	10
28	31	26	29	31	16	

2. At the fifth hockey game of the season at a certain arena, 200 people were selected at random and asked how many of the previous four games they had attended. The results are given in Table 10.7. Test the hypothesis that these 200 observed values can be regarded as a random sample from a binomial distribution; that is, there exists a number θ $(0 < \theta < 1)$ such that the probabilities are as follows:

$$p_0 = (1-\theta)^4, \quad p_1 = 4\theta(1-\theta)^3, \quad p_2 = 6\theta^2(1-\theta)^2,$$
$$p_3 = 4\theta^3(1-\theta), \quad p_4 = \theta^4.$$

Table 10.7 Data for Exercise 2

Number of games previously attended	Number of people
0	33
1	67
2	66
3	15
4	19

3. Consider a genetics problem in which each individual in a certain population must have one of six genotypes, and it is desired to test the null hypothesis H_0 that the probabilities of the six genotypes can be represented in the form specified in Eq. (10.2.2).

a. Suppose that in a random sample of n individuals, the observed numbers of individuals having the six genotypes are $N_1, \dots, N_6$. Find the M.L.E.'s of θ_1 and θ_2 when the null hypothesis H_0 is true.

b. Suppose that in a random sample of 150 individuals, the observed numbers are as follows:

$$N_1 = 2, \quad N_2 = 36, \quad N_3 = 14, \quad N_4 = 36,$$
$$N_5 = 20, \quad N_6 = 42.$$

Determine the value of Q and the corresponding tail area.

4. Consider again the sample consisting of the heights of 500 men given in Exercise 8 of Sec. 10.1. Suppose that before these heights were grouped into the intervals given in that exercise, it was found that for the 500 observed heights in the original sample, the sample mean was $\overline{X}_n = 67.6$ and the sample variance was $S_n^2/n = 1.00$. Test the hypothesis that these observed heights form a random sample from a normal distribution.

5. In a large city, 200 persons were selected at random, and each person was asked how many tickets he purchased that week in the state lottery. The results are given in Table 10.8. Suppose that among the seven persons who had purchased five or more tickets, three persons had purchased exactly five tickets, two persons had purchased six tickets, one person had purchased seven tickets, and one person had purchased 10 tickets. Test the hypothesis that these 200 observations form a random sample from a Poisson distribution.

Table 10.8 Data for Exercise 5

Number of tickets previously purchased	Number of persons
0	52
1	60
2	55
3	18
4	8
5 or more	7

6. Rutherford and Geiger (1910) counted the numbers of alpha particles emitted by a certain mass of polonium during 2608 disjoint time periods, each of which lasted 7.5 seconds. The results are given in Table 10.9. Test the hypothesis that these 2608 observations form a random sample from a Poisson distribution.

Table 10.9 Data for Exercise 6 from Rutherford and Geiger (1910)

Number of particles emitted	Number of time periods
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	45
9	27
10	10
11	4
12	0
13	1
14	1
15 or more	0
Total	2608

7. Test the hypothesis that the 50 observations in Table 10.10 form a random sample from a normal distribution.

Table 10.10 Data for Exercise 7

9.69	8.93	7.61	8.12	−2.74
2.78	7.47	8.46	7.89	5.93
5.21	2.62	0.22	−0.59	8.77
4.07	5.15	8.32	6.01	0.68
9.81	5.61	13.98	10.22	7.89
0.52	6.80	2.90	2.06	11.15
10.22	5.05	6.06	14.51	13.05
9.09	9.20	7.82	8.67	7.52
3.03	5.29	8.68	11.81	7.80
16.80	8.07	0.66	4.01	8.64

8. Test the hypothesis that the 50 observations in Table 10.11 form a random sample from an exponential distribution.

Table 10.11 Data for Exercise 8

0.91	1.22	1.28	0.22	2.33
0.90	0.86	1.45	1.22	0.55
0.16	2.02	1.59	1.73	0.49
1.62	0.56	0.53	0.50	0.24
1.28	0.06	0.19	0.29	0.74
1.16	0.22	0.91	0.04	1.41
3.65	3.41	0.07	0.51	1.27
0.61	0.31	0.22	0.37	0.06
1.75	0.89	0.79	1.28	0.57
0.76	0.05	1.53	1.86	1.28

10.3 Contingency Tables

When each observation in our sample is a bivariate discrete random vector (a pair of discrete random variables), then there is a simple way to test the hypothesis that the two random variables are independent. The test is another form of χ^2 test like the ones used earlier in this chapter.

Independence in Contingency Tables

Example 10.3.1

College Survey. Suppose that 200 students are selected at random from the entire enrollment at a large university, and each student in the sample is classified both according to the curriculum in which he is enrolled and according to his preference for either of two candidates A and B in a forthcoming election. Suppose that the results are as presented in Table 10.12. We might be interested in whether the choices of

Table 10.12 Classification of students by curriculum and candidate preference

	Candidate preferred			
Curriculum	A	B	Undecided	Totals
Engineering and science	24	23	12	59
Humanities and social sciences	24	14	10	48
Fine arts	17	8	13	38
Industrial and public administration	27	19	9	55
Totals	92	64	44	200

curriculum and candidate are independent of each other. To be more precise, suppose that a student is selected at random from the entire enrollment at the university. Independence means that for each i and j, the probability that such a randomly chosen student prefers candidate j and is in curriculum i equals the product of the probability that he prefers candidate j times the probability that he is enrolled in curriculum i. ◀

Tables of data like Table 10.12 are very common and have a special name.

Definition 10.3.1 Contingency Tables. A table in which each observation is classified in two or more ways is called a *contingency table*.

In Table 10.12, only two classifications are considered for each student, namely, the curriculum in which he is enrolled and the candidate he prefers. Such a table is called a *two-way* contingency table.

In general, we shall consider a two-way contingency table containing R rows and C columns. For $i = 1, \dots, R$ and $j = 1, \dots, C$, we shall let p_{ij} denote the probability that an individual selected at random from a given population will be classified in the ith row and the jth column of the table. Furthermore, we shall let p_{i+} denote the marginal probability that the individual will be classified in the ith row of the table and p_{+j} denote the marginal probability that the individual will be classified in the jth column of the table. Thus,

$$p_{i+} = \sum_{j=1}^{C} p_{ij} \quad \text{and} \quad p_{+j} = \sum_{i=1}^{R} p_{ij}.$$

Furthermore, since the sum of the probabilities for all the cells of the table must be 1, we have

$$\sum_{i=1}^{R}\sum_{j=1}^{C} p_{ij} = \sum_{i=1}^{R} p_{i+} = \sum_{j=1}^{C} p_{+j} = 1.$$

Suppose now that a random sample of n individuals is taken from the given population. For $i = 1, \dots, R$, and $j = 1, \dots, C$, we shall let N_{ij} denote the number of individuals who are classified in the ith row and the jth column of the table. Furthermore, we shall let N_{i+} denote the total number of individuals classified in the ith row and N_{+j} denote the total number of individuals classified in the jth column.

Thus,

$$N_{i+} = \sum_{j=1}^{C} N_{ij} \quad \text{and} \quad N_{+j} = \sum_{i=1}^{R} N_{ij}. \tag{10.3.1}$$

Also,

$$\sum_{i=1}^{R}\sum_{j=1}^{C} N_{ij} = \sum_{i=1}^{R} N_{i+} = \sum_{j=1}^{C} N_{+j} = n. \tag{10.3.2}$$

On the basis of these observations, the following hypotheses are to be tested:

$$\begin{aligned} &H_0: \quad p_{ij} = p_{i+}p_{+j} \quad \text{for } i = 1, \ldots, R \text{ and } j = 1, \ldots, C, \\ &H_1: \quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{10.3.3}$$

The χ^2 Test of Independence

The χ^2 tests described in Sec. 10.2 can be applied to the problem of testing the hypotheses (10.3.3). Each individual in the population from which the sample is taken must belong in one of the RC cells of the contingency table. Under the null hypothesis H_0, the unknown probabilities p_{ij} of these cells have been expressed as functions of the unknown parameters p_{i+} and p_{+j}. Since $\sum_{i=1}^{R} p_{i+} = 1$ and $\sum_{j=1}^{C} p_{+j} = 1$, the actual number of unknown parameters to be estimated when H_0 is true is $s = (R-1) + (C-1)$, or $s = R + C - 2$.

For $i = 1, \ldots, R$, and $j = 1, \ldots, C$, let $\hat{E}_{ij}$ denote the M.L.E., when H_0 is true, of the expected number of observations that will be classified in the ith row and the jth column of the table. In this problem, the statistic Q defined by Eq. (10.2.4) will have the following form:

$$Q = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(N_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}. \tag{10.3.4}$$

Furthermore, since the contingency table contains RC cells, and since $s = R + C - 2$ parameters are to be estimated when H_0 is true, it follows that when H_0 is true and $n \to \infty$, the c.d.f. of Q converges to the c.d.f. of the χ^2 distribution for which the number of degrees of freedom is $RC - 1 - s = (R-1)(C-1)$.

Next, we shall consider the form of the estimator $\hat{E}_{ij}$. The expected number of observations in the ith row and the jth column is simply np_{ij}. When H_0 is true, $p_{ij} = p_{i+}p_{+j}$. Therefore, if $\hat{p}_{i+}$ and $\hat{p}_{+j}$ denote the M.L.E.'s of p_{i+} and p_{+j}, then it follows that $\hat{E}_{ij} = n\hat{p}_{i+}\hat{p}_{+j}$. Next, since p_{i+} is the probability that an observation will be classified in the ith row, $\hat{p}_{i+}$ is simply the proportion of observations in the sample that are classified in the ith row; that is, $\hat{p}_{i+} = N_{i+}/n$. Similarly, $\hat{p}_{+j} = N_{+j}/n$, and it follows that

$$\hat{E}_{ij} = n\left(\frac{N_{i+}}{n}\right)\left(\frac{N_{+j}}{n}\right) = \frac{N_{i+}N_{+j}}{n}. \tag{10.3.5}$$

If we substitute this value of $\hat{E}_{ij}$ into Eq. (10.3.4), we can calculate the value of Q from the observed values of N_{ij}. The null hypothesis H_0 should be rejected if $Q \geq c$, where c is an appropriately chosen constant. When H_0 is true, and the sample size n is large, the distribution of Q will be approximately the χ^2 distribution with $(R-1)(C-1)$ degrees of freedom.

Table 10.13 Expected cell counts for Example 10.3.2

Curriculum	Candidate preferred A	B	Undecided	Totals
Engineering and science	27.14	18.88	12.98	59
Humanities and social sciences	22.08	15.36	10.56	48
Fine arts	17.48	12.16	8.36	38
Industrial and public administrations	25.30	17.60	12.10	55
Totals	92	64	44	200

Example 10.3.2

College Survey. Suppose that we wish to test the hypotheses (10.3.3) on the basis of the data in Table 10.12. By using the totals given in the table, we find that $N_{1+} = 59$, $N_{2+} = 48$, $N_{3+} = 38$, and $N_{4+} = 55$, and also $N_{+1} = 92$, $N_{+2} = 64$, and $N_{+3} = 44$. Because $n = 200$, it follows from Eq. (10.3.5) that the 4×3 table of values of $\hat{E}_{ij}$ is as shown in Table 10.13.

The values of N_{ij} given in Table 10.12 can now be compared with the values of $\hat{E}_{ij}$ in Table 10.13. The value of Q defined by Eq. (10.3.4) turns out to be 6.68. Since $R = 4$ and $C = 3$, the corresponding tail area is to be found from a table of the χ^2 distribution with $(R - 1)(C - 1) = 6$ degrees of freedom. Its value is larger than 0.3. Therefore, we would only reject H_0 at level α_0 if $\alpha_0 \geq 0.3$. ◄

Example 10.3.3

Montana Outlook Poll. In Example 10.1.4, we examined the surveyed opinions of Montana residents on their personal financial status. Another question that survey participants were asked was an income range. Table 10.14 gives a cross-tabulation of the answers to both questions. We can use the χ^2 test to test the null hypothesis that income is independent of opinion on personal financial status. Table 10.15 gives the expected counts for each cell of Table 10.14 under the null hypothesis. We can now compute the test statistic $Q = 5.210$ with $(3 - 1) \times (3 - 1) = 4$ degrees of freedom. The p-value associated with this value of Q is 0.266, so we would only reject the null hypothesis at a level α_0 greater than 0.266. ◄

Table 10.14 Responses to two questions from Montana Outlook Poll

Income range	Personal financial status Worse	Same	Better	Total
Under \$20,000	20	15	12	47
\$20,000–\$35,000	24	27	32	83
Over \$35,000	14	22	23	59
Total	58	64	67	189

Table 10.15 Expected cell counts for Table 10.14 under the assumption of independence

	Personal financial status			
Income range	Worse	Same	Better	Total
Under \$20,000	14.42	15.92	16.66	47
\$20,000–\$35,000	25.47	28.11	29.42	83
Over \$35,000	18.11	19.98	20.92	59
Total	58	64	67	189

Summary

We learned how to test the null hypothesis that two discrete random variables are independent based on a random sample of n pairs. First, form a contingency table of the counts for every pair of possible observed values. Then, estimate the two marginal distributions of the two random variables. Under the null hypothesis that the random variables are independent, the expected count for value i of the first variable and value j of the second variable is n times the product of the two estimated marginal probabilities. We then form the χ^2 statistic Q by summing (observed−expected)^2/expected over all of the cells in the contingency table. The degrees of freedom is $(R-1)(C-1)$, where R is the number of rows in the table and C is the number of columns.

Exercises

1. Chase and Dummer (1992) studied the attitudes of school-aged children in Michigan. The children were asked which of the following was most important to them: good grades, athletic ability, or popularity. Additional information about each child was also collected, and Table 10.16 shows the results for 478 children classified by sex and their response to the survey question. Test the null hypothesis that a child's answer to the survey question is independent of his or her sex.

Table 10.16 Data for Exercise 1 from Chase and Dummer (1992)

	Good grades	Athletic ability	Popularity
Boys	117	60	50
Girls	130	30	91

2. Show that the statistic Q defined by Eq. (10.3.4) can be rewritten in the form

$$Q = \left(\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{N_{ij}^2}{\hat{E}_{ij}}\right) - n.$$

3. Show that if $C = 2$, the statistic Q defined by Eq. (10.3.4) can be rewritten in the form

$$Q = \frac{n}{N_{+2}}\left(\sum_{i=1}^{R}\frac{N_{i1}^2}{\hat{E}_{i1}} - N_{+1}\right).$$

4. Suppose that an experiment is carried out to see if there is any relation between a man's age and whether he wears a moustache. Suppose that 100 men, 18 years of age or older, are selected at random, and each man is classified according to whether or not he is between 18 and 30 years of age and also according to whether or not he wears a moustache. The observed numbers are given in Table 10.17. Test the hypothesis that there is no relationship between a man's age and whether he wears a moustache.

Table 10.17 Data for Exercise 4

	Wears a moustache	Does not wear a moustache
Between 18 and 30	12	28
Over 30	8	52

5. Suppose that 300 persons are selected at random from a large population, and each person in the sample is classified according to blood type, O, A, B, or AB, and also according to Rh, positive or negative. The observed numbers are given in Table 10.18. Test the hypothesis that the two classifications of blood types are independent.

Table 10.18 Data for Exercise 5

	O	A	B	AB
Rh positive	82	89	54	19
Rh negative	13	27	7	9

6. Suppose that a store carries two different brands, A and B, of a certain type of breakfast cereal. Suppose that during a one-week period the store noted whether each package of this type of cereal that was purchased was brand A or brand B and also noted whether the purchaser was a man or a woman. (A purchase made by a child or by a man and a woman together was not counted.) Suppose that 44 packages were purchased, and that the results were as shown in Table 10.19. Test the hypothesis that the brand purchased and the sex of the purchaser are independent.

Table 10.19 Data for Exercise 6

	Brand A	Brand B
Men	9	6
Women	13	16

7. Consider a two-way contingency table with three rows and three columns. Suppose that, for $i = 1, 2, 3$ and $j = 1, 2, 3$, the probability p_{ij} that an individual selected at random from a given population will be classified in the ith row and the jth column of Table 10.20.

Table 10.20 Data for Exercise 7

0.15	0.09	0.06
0.15	0.09	0.06
0.20	0.12	0.08

a. Show that the rows and columns of this table are independent by verifying that the values p_{ij} satisfy the null hypothesis H_0 in Eq. (10.3.3).

b. Generate a random sample of 300 observations from the given population using a uniform pseudo-random number generator. Select 300 pseudo-random numbers between 0 and 1 and proceed as follows: Since $p_{11} = 0.15$, classify a pseudo-random number x in the first cell if $x < 0.15$. Since $p_{11} + p_{12} = 0.24$, classify a pseudo-random number x in the second cell if $0.15 \le x < 0.24$. Continue in this way for all nine cells. For example, since the sum of all probabilities except p_{33} is 0.92, a pseudo-random number x will be classified in the lower-right cell of the table if $x \ge 0.92$.

c. Consider the 3×3 table of observed values N_{ij} generated in part (b). Pretend that the probabilities p_{ij} were unknown, and test the hypotheses (10.3.3).

8. If all the students in a class carry out Exercise 7 independently of each other and use different pseudo-random numbers, then the different values of the statistic Q obtained by the different students should form a random sample from the χ^2 distribution with four degrees of freedom. If the values of Q for all the students in the class are available to you, test the hypothesis that these values form such a random sample.

9. Consider a three-way contingency table of size $R \times C \times T$. For $i = 1, \dots, R$, $j = 1, \dots, C$, and $k = 1, \dots, T$, let p_{ijk} denote the probability that an individual selected at random from a given population will fall into the (i, j, k) cell of the table. Let

$$p_{i++} = \sum_{j=1}^{C}\sum_{k=1}^{T} p_{ijk}, \quad p_{+j+} = \sum_{i=1}^{R}\sum_{k=1}^{T} p_{ijk},$$

$$p_{++k} = \sum_{i=1}^{R}\sum_{j=1}^{C} p_{ijk}.$$

On the basis of a random sample of n observations from the given population, construct a test of the following hypotheses:

$$H_0:\ p_{ijk} = p_{i++}p_{+j+}p_{++k} \quad \text{for all values of } i, j, \text{ and } k,$$
$$H_1:\ \text{The hypothesis } H_0 \text{ is not true.}$$

10. Consider again the conditions of Exercise 9. For $i = 1, \dots, R$, and $j = 1, \dots, C$, let

$$p_{ij+} = \sum_{k=1}^{T} p_{ijk}.$$

On the basis of a random sample of n observations from the given population, construct a test of the following hypotheses:

$$H_0:\ p_{ijk} = p_{ij+}p_{++k} \quad \text{for all values of } i, j, \text{ and } k,$$
$$H_1:\ \text{The hypothesis } H_0 \text{ is not true.}$$

10.4 Tests of Homogeneity

Imagine that we select subjects from several different populations, and that we observe a discrete random variable for each subject. We might be interested in whether or not the distribution of that discrete random variable is the same in each population. There is a χ^2 test of this hypothesis that is very similar to the χ^2 test of independence.

Samples from Several Populations

Example 10.4.1

College Survey. Consider again the problem described in Example 10.3.1. There we assumed that a random sample of 200 students was drawn from the entire enrollment at a large university and classified in a contingency table according to the curriculum in which he is enrolled and according to his preference for either of two political candidates A and B. The resulting table appears in Table 10.12.

Suppose, now, that instead of sampling 200 students at random, we had actually sampled separately from each of the four curricula. That is, suppose that we had sampled 59 students at random from those enrolled in engineering and science along with 48 students selected at random from those enrolled in humanities and social sciences and 38 from those enrolled in fine arts and 55 from those enrolled in industrial and public administration. After the students are sampled, those in each curriculum are then classified according to whether they prefer candidate A or B, or are undecided. Suppose that the responses within each curriculum are the same as those reported in Table 10.12.

We might still be interested in investigating whether there is a relationship between the curriculum in which a student is enrolled and the candidate he prefers. This time, we might word the question of interest as follows: Are the distributions of candidate preferences within the different curricula the same or do the students in different curricula have different distributions of preferences among the candidates? ◀

In Example 10.4.1, we are assuming that we have obtained a table of values identical to Table 10.12; we are assuming now that this table was obtained by taking four different random samples from the different populations of students defined by the four rows of the table. This is in contrast to Example 10.3.1, in which we assumed that all students were drawn from one population and then classified according to the values of two variables: preference and curriculum. In the present context, we are interested in testing the hypothesis that, in all four populations, the same proportion of students prefers candidate A, the same proportion prefers candidate B, and the same proportion is undecided.

In general, we shall consider a problem in which random samples are taken from R different populations, and each observation in each sample can be classified as one of C different types. Thus, the data obtained from the R samples can be represented in an $R \times C$ table. For $i = 1, \dots, R$, and $j = 1, \dots, C$, we shall let p_{ij} denote the probability that an observation chosen at random from the ith population will be of type j. Thus,

$$\sum_{j=1}^{C} p_{ij} = 1 \quad \text{for } i = 1, \dots, R.$$

The hypotheses to be tested are as follows:

$$
\begin{aligned}
&H_0: \quad p_{1j} = p_{2j} = \cdots = p_{Rj} \quad \text{for } j = 1, \ldots, C, \\
&H_1: \quad \text{The hypothesis } H_0 \text{ is not true.}
\end{aligned}
\tag{10.4.1}
$$

The null hypothesis H_0 in (10.4.1) states that all the distributions from which the R different samples are drawn are actually alike, that is, that the R distributions are identical. If the null hypothesis in (10.4.1) were true, then combining the R populations would produce one homogeneous population with regard to the distribution of the random variables we are studying. For this reason, a test of the hypotheses (10.4.1) is called a *test of homogeneity* of the R distributions.

For $i = 1, \ldots, R$, we shall let N_{i+} denote the number of observations in the random sample from the ith population; for $j = 1, \ldots, C$, we shall let N_{ij} denote the number of observations in this random sample that are of type j. Thus,

$$
\sum_{j=1}^{C} N_{ij} = N_{i+} \quad \text{for } i = 1, \ldots, R.
$$

Furthermore, if we let n denote the total number of observations in all R samples and N_{+j} denote the total number of observations of type j in the R samples, then all the relations in Eqs. (10.3.1) and (10.3.2) will again be satisfied.

The χ^2 Test of Homogeneity

We shall now develop a test procedure for the hypotheses (10.4.1). Suppose for the moment that the probabilities p_{ij} are known, and consider the following statistic calculated from the observations in the ith random sample:

$$
\sum_{j=1}^{C} \frac{(N_{ij} - N_{i+}p_{ij})^2}{N_{i+}p_{ij}}.
$$

This statistic is just the standard χ^2 statistic, introduced in Eq. (10.1.2), for the random sample of N_{i+} observations from the ith population. Therefore, when the sample size N_{i+} is large, the distribution of this statistic will be approximately the χ^2 distribution with $C - 1$ degrees of freedom.

If we now sum this statistic over the R different samples, we obtain the following statistic:

$$
\sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(N_{ij} - N_{i+}p_{ij})^2}{N_{i+}p_{ij}}. \tag{10.4.2}
$$

Since the observations in the R samples are drawn independently, the distribution of the statistic (10.4.2) will be the distribution of the sum of R independent random variables, each of which has approximately the χ^2 distribution with $C - 1$ degrees of freedom. Hence, the distribution of the statistic (10.4.2) will be approximately the χ^2 distribution with $R(C - 1)$ degrees of freedom.

Since the probabilities p_{ij} are not actually known, their values must be estimated from the observed numbers in the R random samples. When the null hypothesis H_0 is true, the R random samples are actually drawn from the same distribution. Therefore, the M.L.E. of the probability that an observation in each of these samples will be of type j is simply the proportion of all the observations in the R samples that are of type j. In other words, the M.L.E. of p_{ij} is the same for all values of i ($i = 1, \ldots, R$), and this estimator is $\hat{p}_{ij} = N_{+j}/n$. When this M.L.E. is substituted into (10.4.2), we

obtain the statistic

$$Q = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(N_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}, \tag{10.4.3}$$

where

$$\hat{E}_{ij} = \frac{N_{i+}N_{+j}}{n}. \tag{10.4.4}$$

It can be seen that Eqs. (10.4.3) and (10.4.4) are precisely the same as Eqs. (10.3.4) and (10.3.5). Thus, the statistic Q to be used for the test of homogeneity in this section is precisely the same as the statistic Q to be used for the test of independence in Sec. 10.3. We shall now show that the number of degrees of freedom is also precisely the same for the test of homogeneity as for the test of independence.

Because the distributions of the R populations are alike when H_0 is true, and because $\sum_{j=1}^{C} p_{ij} = 1$ for this common distribution, we have estimated $C - 1$ parameters in this problem. Therefore, the statistic Q will have approximately the χ^2 distribution with $R(C - 1) - (C - 1) = (R - 1)(C - 1)$ degrees of freedom. This number is the same as that found in Sec. 10.3.

In summary, consider Table 10.12 again. The statistical analysis of this table will be the same for either of the following two procedures: The 200 observations are drawn as a single random sample from the entire enrollment of the university, and a test of independence is carried out; or the 200 observations are drawn as separate random samples from four different groups of students, and a test of homogeneity is carried out. In either case, in a problem of this type with R rows and C columns, we should calculate the statistic Q defined by Eqs. (10.4.3) and (10.4.4), and we should assume that its distribution when H_0 is true will be approximately the χ^2 distribution with $(R - 1)(C - 1)$ degrees of freedom.

Note: Why the two χ^2 tests look so similar. The reason that the same calculation is appropriate for both the χ^2 test of independence and the χ^2 test of homogeneity is the following: First, consider the situation of Sec. 10.3, in which one sample is drawn and the random variables corresponding to rows and columns are measured. Independence of the row and column variables is equivalent to the conditional distribution of the column variable given a value of the row variable being the same for every value of the row variable. Hence, the test of independence tests that the conditional distributions of the column variable are the same for each value of the row variable. Next, think of the row variable as defining subpopulations (for example, different curricula in Table 10.12). The conditional distributions of the column variable given each value of the row variable are the distributions of the column variable within each subpopulation. The test of homogeneity tests that the distributions within the subpopulations are the same if the samples had been drawn separately from each subpopulation rather than drawn at random from the entire population.

Comparing Two or More Proportions

Example 10.4.2

Television Survey. Suppose that independent samples are drawn from adults in several cities. Each sampled person is asked whether or not they watched a particular television program. Suppose that we want to test the null hypothesis H_0 that the proportion of adults who watched a certain television program was the same in each of the cities. To be specific, suppose that there are R different cities ($R \geq 2$). Suppose

Table 10.21 Form of table for comparing two or more proportions

City	Watched program	Did not watch	Sample size
1	N_{11}	N_{12}	N_{1+}
2	N_{21}	N_{22}	N_{2+}
$\vdots$			
R	N_{R1}	N_{R2}	N_{R+}

that for $i = 1, \ldots, R$, a random sample of N_{i+} adults is selected from city i, the number in the sample who watched the program is N_{i1}, and the number who did not watch the program is $N_{i2} = N_{i+} - N_{i1}$. These data can be presented in an $R \times 2$ table such as Table 10.21. The hypotheses to be tested will have the same form as the hypotheses (10.4.1). Hence, when the null hypothesis H_0 is true, that is, when the proportion of adults who watched the program is the same in all R cities, the statistic Q defined by Eqs. (10.4.3) and (10.4.4) will have approximately the χ^2 distribution with $R - 1$ degrees of freedom. ◄

The reasoning in Example 10.4.2 extends to other problems in which we wish to compare a collection of proportions.

Example 10.4.3

A Clinical Trial. The data in Table 2.1 (see Example 2.1.4 on page 57) are the numbers of subjects in four different treatment groups in a clinical trial together with the numbers who did or did not relapse after treatment. We might wish to test the null hypothesis that the probability of no relapse is the same in all four treatment groups. We can easily compute the statistic Q in Eq. (10.4.3) to be 10.80. This is the 0.987 quantile of the χ^2 distribution with three degrees of freedom. That is, the p-value is 0.013, and the null hypothesis of equal probabilities would be rejected at every level $\alpha_0 \geq 0.013$. ◄

Correlated 2 × 2 Tables

We shall now describe a type of problem in which the use of the χ^2 test of homogeneity would not be appropriate. Suppose that 100 persons were selected at random in a certain city, and that each person was asked whether she thought the service provided by the fire department in the city was satisfactory. Shortly after this survey was carried out, a large fire occurred in the city. Suppose that after this fire, the same 100 persons were again asked whether they thought that the service provided by the fire department was satisfactory. The results are presented in Table 10.22.

Table 10.22 has the same general appearance as other tables we have been considering in this section. However, it would not be appropriate to carry out a χ^2 test of homogeneity for this table, because the observations taken before the fire and the observations taken after the fire are not independent. Although the total number of observations in Table 10.22 is 200, only 100 independently chosen persons were questioned in the surveys. It is reasonable to believe that a particular person's

Table 10.22 Correlated 2×2 table

	Satisfactory	Unsatisfactory
Before the fire	80	20
After the fire	72	28

Table 10.23 2×2 table for correlated responses

	After the fire	
Before the fire	Satisfactory	Unsatisfactory
Satisfactory	70	10
Unsatisfactory	2	18

opinion before the fire and her opinion after the fire are dependent. For this reason, Table 10.22 is called a correlated 2×2 table.

The proper way to display the opinions of the 100 persons in the random sample is shown in Table 10.23. It is not possible to construct Table 10.23 from the data in Table 10.22 alone. The entries in Table 10.22 are simply the marginal totals of Table 10.23. However, in order to construct Table 10.23, it is necessary to go back to the original data and, for each person in the sample, to consider her opinion before the fire and her opinion after the fire.

Furthermore, it usually is not appropriate to carry out either a χ^2 test of independence or a χ^2 test of homogeneity for Table 10.23, because the hypotheses that are tested by either of these procedures usually are not those in which a researcher would be interested in this type of problem. In fact, in this problem a researcher would basically be interested in the answers to one or both of the following two questions: First, what proportion of the persons in the city changed their opinions about the fire department after the fire occurred? Second, among those persons in the city who did change their opinions after the fire, were the changes predominantly in one direction rather than the other?

Table 10.23 provides information pertaining to both these questions. According to Table 10.23, the number of persons in the sample who changed their opinions after the fire was $10 + 2 = 12$. Furthermore, among the 12 persons who did change their opinions, the opinions of 10 of them were changed from satisfactory to unsatisfactory and the opinions of two of them were changed from unsatisfactory to satisfactory. On the basis of these statistics, it is possible to make inferences about the corresponding proportions for the entire population of the city.

In this example, the M.L.E. $\hat{\theta}$ of the proportion of the population who changed their opinions after the fire is 0.12. Also, among those who did change their opinions, the M.L.E. $\hat{p}_{12}$ of the proportion who changed from satisfactory to unsatisfactory is 5/6. Of course, if $\hat{\theta}$ is very small in a particular problem, then there is little interest in the value of $\hat{p}_{12}$.

Summary

When we sample discrete random variables from several populations, we might be interested in the null hypothesis that the distribution of the random variables is the same in all populations. We can perform a χ^2 test of this null hypothesis as follows: Create a new variable with values equal to the names of the different populations. Next, pretend as if each observation consists of the original discrete random variable together with the new "population name" variable. Finally, compute the χ^2 test statistic Q from Sec. 10.3 with the same degrees of freedom. For the type of data considered in this section, the "population name" for each observation is known before sampling begins, and hence it is not a random variable. Whether the population name is known ahead of time or is observed as part of the sampled data (as in Sec. 10.3), the mechanics of the χ^2 test are the same.

Exercises

1. The survey of Chase and Dummer (1992) discussed in Exercise 1 of Sec. 10.3 was actually collected by sampling from three subpopulations according to the locations of the schools: rural, suburban, and urban. Table 10.24 shows the responses to the survey question classified by school location. Test the null hypothesis that the distribution of responses is the same in all three types of school location.

Table 10.24 Data for Exercise 1 from Chase and Dummer (1992)

	Good grades	Athletic ability	Popularity
Rural	57	42	50
Suburban	87	22	42
Urban	103	26	49

2. An examination was given to 500 high school seniors in each of two large cities, and their grades were recorded as low, medium, or high. The results are given in Table 10.25. Test the hypothesis that the distributions of scores among seniors in the two cities are the same.

Table 10.25 Data for Exercise 2

	Low	Medium	High
City A	103	145	252
City B	140	136	224

3. Every Tuesday afternoon during the school year, a certain university brought in a visiting speaker to present a lecture on some topic of current interest. On the day after the fourth lecture of the year, random samples of 70 freshmen, 70 sophomores, 60 juniors, and 50 seniors were selected from the student body at the university, and each of these students was asked how many of the four lectures she had attended. The results are given in Table 10.26. Test the hypothesis that freshmen, sophomores, juniors, and seniors at the university attended the lectures with equal frequency.

Table 10.26 Data for Exercise 3

	Number of lectures attended				
	0	1	2	3	4
Freshmen	10	16	27	6	11
Sophomores	14	19	20	4	13
Juniors	15	15	17	4	9
Seniors	19	8	6	5	12

4. Suppose that five persons shoot at a target. Suppose also that for $i = 1, \dots, 5$, person i shoots n_i times and hits the target y_i times, and that the values of n_i and y_i are as given in Table 10.27. Test the hypothesis that the five persons are equally good marksmen.

Table 10.27 Data for Exercise 4

i	n_i	y_i
1	17	8
2	16	4
3	10	7
4	24	13
5	16	10

5. A manufacturing plant has preliminary contracts with three different suppliers of machines. Each supplier delivered 15 machines, which were used in the plant for four months in preliminary production. It turned out that one of the machines from supplier 1 was defective, seven of the machines from supplier 2 were defective, and seven of the machines from supplier 3 were defective. The plant statistician decided to test the null hypothesis H_0 that the three suppliers provided the same quality. Therefore, he set up Table 10.28 and carried out a χ^2 test. By summing the values in the bottom row of Table 10.28, he found that the value of the χ^2 statistic was 24/5 with two degrees of freedom. He then found from a table of the χ^2 distribution that H_0 should be accepted when the level of significance is 0.05. Criticize this procedure and provide a meaningful analysis of the observed data.

Table 10.28 Data for Exercise 5

	Supplier		
	1	2	3
Number of defectives N_i	1	7	7
Expected number of defectives E_i under H_0	5	5	5
$\frac{(N_i - E_i)^2}{E_i}$	$\frac{16}{5}$	$\frac{4}{5}$	$\frac{4}{5}$

6. Suppose that 100 students in a physical education class shoot at a target with a bow and arrow, and 27 students hit the target. These 100 students are then given a demonstration on the proper technique for shooting with the bow and arrow. After the demonstration, they again shoot at the target. This time 35 students hit the target. What additional information, if any, is needed in order to investigate the hypothesis that the demonstration was helpful?

7. As people entered a certain meeting, n persons were selected at random, and each was asked either to name one of two political candidates she favored in a forthcoming election or to say "undecided" if she had no real preference. During the meeting, the people heard a speech on behalf of one of the candidates. After the meeting, each of the same n persons was again asked to express her opinion. Describe a method for evaluating the effectiveness of the speaker.

10.5 Simpson's Paradox

When tabulating discrete data, we need to be careful about aggregating groups. Suppose that a survey has two questions. If we construct a single table of responses to the two questions that includes both men and women, we might get a very different picture than if we construct separate tables for the responses of men and women.

An Example of the Paradox

Example 10.5.1

Comparing Treatments in an Aggregated Table. Suppose that an experiment is carried out in order to compare a new treatment for a particular disease with the standard treatment for the disease. In the experiment, 80 subjects suffering from the disease are treated, 40 subjects receiving the new treatment and 40 receiving the standard treatment. After a certain period of time, it is observed how many of the subjects in each group have improved and how many have not. Suppose that the overall results for all 80 patients are as shown in Table 10.29.

According to this table, 20 of the 40 subjects who received the new treatment improved, and 24 of the 40 subjects who received the standard treatment improved. Thus, 50 percent of the subjects improved under the new treatment, whereas 60 percent improved under the standard treatment. On the basis of these results, the new treatment appears inferior to the standard treatment. ◄

Table 10.29 Results of experiment comparing two treatments

All patients	Improved	Not improved	Percent improved
New treatment	20	20	50
Standard treatment	24	16	60

Table 10.30 Table 10.29 disaggregated by sex

Men only	Improved	Not improved	Percent improved
New treatment	12	18	40
Standard treatment	3	7	30
Women only			
New treatment	8	2	80
Standard treatment	21	9	70

Many contingency tables, such as Table 10.29, summarize the results of a study in only one of several possible ways. The next example looks at the same data from a different point of view and draws a different conclusion.

Example 10.5.2

Comparing Treatments in an Disaggregated Table. In order to investigate more carefully the efficacy of the new treatment in Example 10.5.1, we might compare it with the standard treatment just for the men in the sample and, separately, just for the women in the sample. The results in Table 10.29 can thus be partitioned into two tables, one pertaining just to men and the other just to women. This process of splitting the overall data into disjoint components pertaining to different subgroups of the population is called *disaggregation*.

Suppose that when the values in Table 10.29 are disaggregated by considering the men and the women separately, the results are as shown in Table 10.30. It can be verified that when the data in these separate tables are combined, or *aggregated*, we again obtain Table 10.29. However, Table 10.30 contains a big surprise because the new treatment appears to be superior to the standard treatment both for men and for women. Specifically, 40 percent of the men (12 out of 30) who received the new treatment improved, but only 30 percent of the men (3 out of 10) who received the standard treatment improved. Furthermore, 80 percent of the women (8 out of 10) who received the new treatment improved, but only 70 percent of the women (21 out of 30) who received the standard treatment improved. ◀

Tables 10.29 and 10.30 together yield somewhat anomalous results. According to Table 10.30, the new treatment is superior to the standard treatment both for men and for women, but according to Table 10.29, the new treatment is inferior to the

standard treatment when all the subjects are aggregated. This type of result is known as *Simpson's paradox*.

It should be emphasized that Simpson's paradox is *not* a phenomenon that occurs because we are working with small samples. The small numbers in Tables 10.29 and 10.30 were used merely for convenience in this explanation. Each of the entries in these tables could be multiplied by 1000 or by 1,000,000 without changing the results.

The Paradox Explained

Of course, Simpson's paradox is not actually a paradox; it is merely a result that is surprising and puzzling to someone who has not seen or thought about it before. It can be seen from Table 10.30 that in the example we are considering, women have a higher rate of improvement from the disease than men have, regardless of which treatment they receive. Furthermore, most of the women in the sample received the standard treatment while most of the men received the new treatment. Specifically, among the 40 men in the sample, 30 received the new treatment, and only 10 received the standard treatment, whereas among the 40 women in the sample, these numbers are reversed.

The new treatment looks bad in the aggregated table because most of the people who weren't going to respond well to either treatment got the new treatment while most of the people who were going to respond well to either treatment got the standard treatment. Even though the numbers of men and women in the experiment were equal, a high proportion of the women and a low proportion of the men received the standard treatment. Since women have a much higher rate of improvement than men, it is found in the aggregated Table 10.29 that the standard treatment manifests a higher overall rate of improvement than does the new treatment.

Simpson's paradox demonstrates dramatically the dangers in making inferences from an aggregated table like Table 10.29. To make sure that Simpson's paradox cannot occur in an experiment like the one just described, the proportions of men and women among the subjects who receive the new treatment must be the same, or approximately the same, as the proportions of men and women among the subjects who receive the standard treatment. It is *not* necessary that there be equal numbers of men and women in the sample.

We can express Simpson's paradox in probability terms. Let A denote the event that a subject chosen for the experiment will be a man, and let A^c denote the event that the subject will be a woman. Also, let B denote the event that a subject will receive the new treatment, and let B^c denote the event that the subject will receive the standard treatment. Finally, let I denote the event that a subject will improve. Simpson's paradox then reflects the fact that it is possible for all three of the following inequalities to hold simultaneously:

$$\begin{aligned} \Pr(I|A \cap B) &> \Pr(I|A \cap B^c), \\ \Pr(I|A^c \cap B) &> \Pr(I|A^c \cap B^c), \\ \Pr(I|B) &< \Pr(I|B^c). \end{aligned} \tag{10.5.1}$$

The discussion that we have just given in regard to the prevention of Simpson's paradox can be expressed as follows: If $\Pr(A|B) = \Pr(A|B^c)$, then it is not possible for all three inequalities in (10.5.1) to hold (see Exercise 5). Similarly, if $\Pr(B|A) = \Pr(B|A^c)$, then it is not possible for all three inequalities in (10.5.1) to hold (see Exercise 3).

The possibility of Simpson's paradox lurks within every contingency table. Even though we might take care to design a particular experiment so that Simpson's

paradox cannot occur when we disaggregate with respect to men and women, it is always possible that there is some other variable, such as the age of the subject or the intensity and the stage of the disease, with respect to which disaggregation would lead us to a conclusion directly opposite to that indicated by the aggregated table. Once an experiment is designed to prevent Simpson's paradox with respect to disaggregations that can be identified in advance, subjects are generally assiged randomly to the possible treatments in the hopes of minimizing the chance that Simpson's paradox will arise with respect to an unforeseen disaggrageation.

Example 10.5.3 Comparing Treatments in an Aggregated Table. In the example of this section, it would be sensible to assign 20 men and 20 women to each of the two treatments. Which 20 men and which 20 women get assigned to each treatment would be determined by randomization in order to minimize the chance of an unforeseen occurrence of Simpson's paradox.

If there were other information, such as severity of disease, that were available at the start of the experiment, the groups of men and women should each be partitioned according to that additional information before being randomly assigned to the treatments. For example, suppose that 12 men and 8 women have more severe cases of the disease before the experiment begins. We should then assign 6 of the men and 4 of the women with more severe cases to each tretment. We should also assign 4 of the men and 6 of the women with less severe cases to each treatment. This balances the factors (sex, severity, and treatment) that are expected to affect the experimental outcome. If there is another unforeseen factor that will affect the outcome, it is still possible, but unlikely, that the random assignment described above will allow Simpson's paradox to arise with regard to that one factor. If there are dozens of additional important factors, some degree of imbalance will be inevitable even with a randomized assignment. ◄

Summary

Simpson's paradox occurs when the relationship between the two categorical variables in every part of a disaggregated table is the opposite of the relationship between those same two variables in the aggregated table.

Exercises

1. Consider two populations I and II. Suppose that 80 percent of the men and 30 percent of the women in population I have a certain characteristic, and that only 60 percent of the men and 10 percent of the women in population II have the characteristic. Explain how, under these conditions, it might be true that the proportion of population II having the characteristic is larger than the proportion of population I having the characteristic.

2. Suppose that A and B are events such that $0 < \Pr(A) < 1$ and $0 < \Pr(B) < 1$. Show that $\Pr(A|B) = \Pr(A|B^c)$ if and only if $\Pr(B|A) = \Pr(B|A^c)$.

3. Show that all three inequalities in (10.5.1) cannot hold if $\Pr(B|A) = \Pr(B|A^c)$.

4. Suppose that each adult subject in an experiment is given either treatment I or treatment II. Prove that the proportion of men among the subjects who receive treatment I is equal to the proportion of men among the subjects who receive treatment II if and only if the proportion of all men in the experiment who receive treatment I is equal to the proportion of all women who receive treatment I.

5. Show that all three inequalities in (10.5.1) cannot hold if $\Pr(A|B) = \Pr(A|B^c)$.

6. It was believed that a certain university was discriminating against women in its admissions policy because 30 percent of all the male applicants to the university were

admitted, whereas only 20 percent of all the female applicants were admitted. In order to determine which of the five colleges in the university were most responsible for this discrimination, the admissions rates for each college were analyzed separately. Surprisingly, it was found that in each college the proportion of female applicants who were admitted to the college was actually larger than the proportion of male applicants who were admitted. Discuss and explain this result.

7. In an experiment involving 800 subjects, each subject received either treatment I or treatment II, and each subject was classified into one of the following four categories: older males, younger males, older females, and younger females. At the end of the experiment, it was determined for each subject whether the treatment that the subject had received was helpful or not. The results for each of the four categories of subjects are given in Table 10.31.

a. Show that treatment II is more helpful than treatment I within each of the four categories of subjects.

b. Show that if these four categories are aggregated into only the two categories, older subjects and younger subjects, then treatment I is more helpful than treatment II within each of these categories.

c. Show that if the two categories in part (b) are aggregated into a single category containing all 800 subjects, then treatment II again appears to be more helpful than treatment I.

Table 10.31 Data for Exercise 7

Older males	Helpful	Not
Treatment I	120	120
Treatment II	20	10
Younger males		
Treatment I	60	20
Treatment II	40	10
Older females		
Treatment I	10	50
Treatment II	20	50
Younger females		
Treatment I	10	10
Treatment II	160	90

⋆ 10.6 Kolmogorov-Smirnov Tests

In Sec. 10.1, we used the χ^2 test to test the null hypothesis that a random sample came from a particular continuous distribution against the alternative hypothesis that the sample did not come from that distribution. A more suitable test for these hypotheses is introduced in this section. This test can also be extended to test the null hypothesis that two independent samples came from the same distribution against the alternative hypothesis that they came from two different distributions.

The Sample Distribution Function

Example 10.6.1

Failure Times of Ball Bearings. In Example 10.1.6, we used a χ^2 goodness-of-fit test to test the null hypothesis that the log-failure times of ball bearings came from the normal distribution with mean 3.912 and variance 0.25. That test required us to choose a somewhat arbitrary partition of the real line in order to convert the log-failure times into count data. Is there a test procedure for such problems that does not require an arbitrary aggregation into intervals that may have no physical meaning in the application? ◄

The first step in trying to answer the question in Example 10.6.1 is to construct an estimator of the distribution of the random sample that does not rely on the assumption that the distribution was normal. Suppose that the random variables $X_1, \ldots, X_n$

form a random sample from some continuous distribution, and let $x_1, \ldots, x_n$ denote the observed values of $X_1, \ldots, X_n$. Since the observations come from a continuous distribution, there is probability 0 that any two of the observed values $x_1, \ldots, x_n$ will be equal. Therefore, we shall assume for simplicity that all n values are different. We shall consider now a function $F_n(x)$, which is constructed from the values $x_1, \ldots, x_n$ and will serve as an estimate of the c.d.f. from which the sample was drawn.

Definition 10.6.1 Sample (Empirical) Distribution Function. Let $x_1, \ldots, x_n$ be the observed values of a random sample $X_1, \ldots, X_n$. For each number x $(-\infty < x < \infty)$, define the value $F_n(x)$ as the proportion of observed values in the sample that are less than or equal to x. In other words, if exactly k of the observed values in the sample are less than or equal to x, then $F_n(x) = k/n$. The function $F_n(x)$ defined in this way is called the *sample distribution function*, or simply the *sample c.d.f.* Sometimes $F_n(x)$ is called the *empirical c.d.f.*

The sample c.d.f. for the data discussed in Example 10.6.1 appears in Fig. 10.1 together with the hypothesized normal c.d.f. mentioned in that example.

In general, the sample c.d.f. $F_n(x)$ can be regarded as the c.d.f. of a discrete distribution that assigns probability $1/n$ to each of the n values $x_1, \ldots, x_n$. Thus, $F_n(x)$ will be a step function with a jump of magnitude $1/n$ at each point x_i $(i = 1, \ldots, n)$. If we let $y_1 < y_2 < \cdots < y_n$ denote the values of the order statistics of the sample, as defined in Definition 7.8.2, then $F_n(x) = 0$ for $x < y_1$; $F_n(x)$ jumps to the value $1/n$ at $x = y_1$ and remains at $1/n$ for $y_1 \le x < y_2$; $F_n(x)$ jumps to the value $2/n$ at $x = y_2$ and remains at $2/n$ for $y_2 \le x < y_3$; and so on.

Now let $F(x)$ denote the c.d.f. of the distribution from which the random sample $X_1, \ldots, X_n$ was drawn. For each given number x $(-\infty < x < \infty)$, the probability that any particular observation X_i will be less than or equal to x is $F(x)$. Therefore, it follows from the law of large numbers that as $n \to \infty$, the proportion $F_n(x)$ of observations in the sample that are less than or equal to x will converge in probability to $F(x)$. In symbols,

$$F_n(x) \xrightarrow{p} F(x) \quad \text{for } -\infty < x < \infty. \tag{10.6.1}$$

The relation (10.6.1) expresses the fact that at each point x, the sample c.d.f. $F_n(x)$ will converge to the actual c.d.f. $F(x)$ of the distribution from which the random sample was taken. A collection of sample c.d.f.'s is sketched in Fig. 10.2 for a few different sized samples from the the same distribution.

Figure 10.1 Sample c.d.f. of log-failure times of ball bearings together with the c.d.f. of the normal distribution with mean 3.912 and variance 0.25.

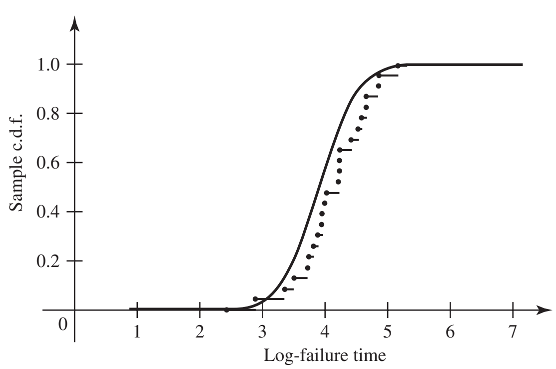


Figure 10.2 The sample c.d.f. $F_n(x)$ for $n = 4, 8, 16$.

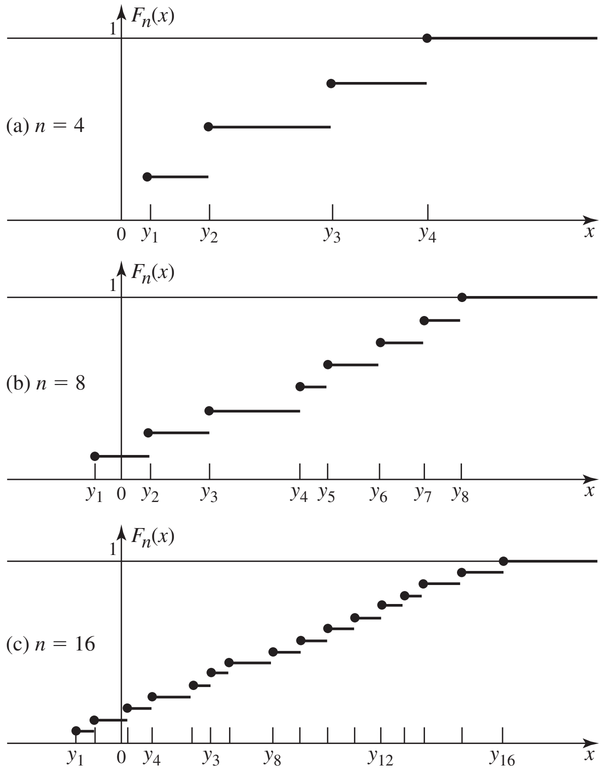
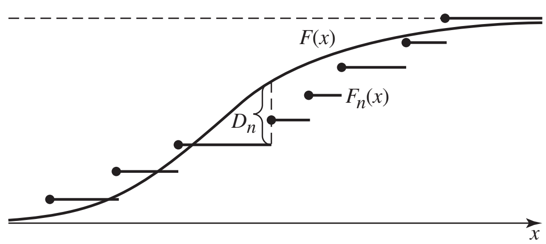


Figure 10.3 The value of D_n.



An even stronger result, known as the Glivenko-Cantelli lemma, states that $F_n(x)$ will converge to $F(x)$ uniformly over all values of x. The proof is beyond the scope of this book.

Theorem 10.6.1 Glivenko-Cantelli Lemma. Let F_n be the sample c.d.f. from an i.i.d. sample $X_1, \ldots, X_n$ from the c.d.f. F. Define

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F(x)|. \tag{10.6.2}$$

Then $D_n \xrightarrow{p} 0$. ■

A value of D_n is illustrated in Fig. 10.3 for a typical example. Before the values of $X_1, \ldots, X_n$ have been observed, the value of D_n is a random variable.

Theorem 10.6.1 implies that when the sample size n is large, the sample c.d.f. $F_n(x)$ is quite likely to be close to the c.d.f. $F(x)$ over the entire real line. In this sense,

when the c.d.f. $F(x)$ is unknown, the sample c.d.f. $F_n(x)$ can be considered to be an estimator of $F(x)$. In another sense, however, $F_n(x)$ is not a very reasonable estimator of $F(x)$. As we explained earlier, $F_n(x)$ will be the c.d.f. of a discrete distribution that is concentrated on n points, whereas we are assuming in this section that the unknown c.d.f. $F(x)$ is the c.d.f. of a continuous distribution. Some type of smoothed version of $F_n(x)$, from which the jumps have been removed, might yield a reasonable estimator of $F(x)$, but we shall not pursue this topic further here.

The Kolmogorov-Smirnov Test of a Simple Hypothesis

Suppose now that we wish to test the simple null hypothesis that the unknown c.d.f. $F(x)$ is actually a particular continuous c.d.f. $F^*(x)$ against the general alternative that the actual c.d.f. is not $F^*(x)$. In other words, suppose that we wish to test the following hypotheses:

$$\begin{aligned} &H_0: \quad F(x) = F^*(x) \quad \text{for } -\infty < x < \infty, \\ &H_1: \quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{10.6.3}$$

This problem is a nonparametric problem because the unknown distribution from which the random sample is taken might be any continuous distribution.

In Sec. 10.1, we described how the χ^2 test of goodness-of-fit can be used to test hypotheses having the form (10.6.3). That test, however, requires grouping the observations into a finite number of intervals in an arbitrary manner. We shall now describe a test of the hypotheses (10.6.3) that does not require such grouping.

As before, we shall let $F_n(x)$ denote the sample c.d.f. Also, we shall now let D_n^* denote the following statistic:

$$D_n^* = \sup_{-\infty<x<\infty} |F_n(x) - F^*(x)|. \tag{10.6.4}$$

In other words, D_n^* is the maximum difference between the sample c.d.f. $F_n(x)$ and the hypothesized c.d.f. $F^*(x)$. When the null hypothesis H_0 in (10.6.3) is true, the probability distribution of D_n^* will be a certain distribution that is the same for every possible continuous c.d.f. $F^*(x)$ and does not depend on the particular c.d.f. $F^*(x)$ being studied in a specific problem. (See Exercise 13.) Tables of this distribution, for various values of the sample size n, have been developed and are presented in many published collections of statistical tables.

It follows from the Glivenko-Cantelli lemma that the value of D_n^* will tend to be small if the null hypothesis H_0 is true, and D_n^* will tend to be larger if the actual c.d.f. $F(x)$ is different from $F^*(x)$. Therefore, a reasonable test procedure for the hypotheses (10.6.3) is to reject H_0 if $n^{1/2}D_n^* > c$, where c is an appropriate constant.

It is convenient to express the test procedure in terms of $n^{1/2}D_n^*$ rather than simply D_n^*, because of the following result, which was established in the 1930s by A. N. Kolmogorov and N. V. Smirnov.

Theorem 10.6.2 If the null hypothesis H_0 is true, then for each given value $t > 0$,

$$\lim_{n\to\infty} \Pr(n^{1/2}D_n^* \le t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2}. \tag{10.6.5}$$

Thus, if the null hypothesis H_0 is true, then as $n \to \infty$, the c.d.f. of $n^{1/2}D_n^*$ will converge to the c.d.f. given by the infinite series on the right side of Eq. (10.6.5). For each value of $t > 0$, we shall let $H(t)$ denote the value on the right side of Eq. (10.6.5). The values of $H(t)$ are given in Table 10.32.

Table 10.32 The c.d.f. H in Eq. (10.6.5)

t	$H(t)$	t	$H(t)$
0.30	0.0000	1.20	0.8878
0.35	0.0003	1.25	0.9121
0.40	0.0028	1.30	0.9319
0.45	0.0126	1.35	0.9478
0.50	0.0361	1.40	0.9603
0.55	0.0772	1.45	0.9702
0.60	0.1357	1.50	0.9778
0.65	0.2080	1.60	0.9880
0.70	0.2888	1.70	0.9938
0.75	0.3728	1.80	0.9969
0.80	0.4559	1.90	0.9985
0.85	0.5347	2.00	0.9993
0.90	0.6073	2.10	0.9997
0.95	0.6725	2.20	0.9999
1.00	0.7300	2.30	0.9999
1.05	0.7798	2.40	1.0000
1.10	0.8223	2.50	1.0000
1.15	0.8580		

Definition 10.6.2 Kolmogorov-Smirnov test. A test procedure that rejects H_0 when $n^{1/2}D_n^* \geq c$ is called a *Kolmogorov-Smirnov test*.

It follows from Eq. (10.6.5) that when the sample size n is large, the constant c can be chosen from Table 10.32 to achieve, at least approximately, any specified level of significance α_0 $(0 < \alpha_0 < 1)$. In fact, we should choose c to be the $1 - \alpha_0$ quantile $H^{-1}(1 - \alpha_0)$ of the distribution H. For example, by examining Table 10.32, we see that $H(1.36) \approx 0.95$, so $H^{-1}(1 - 0.05) = 1.36$. Therefore, if the null hypothesis H_0 is true, then $\Pr(n^{1/2}D_n^* \geq 1.36) = 0.05$. It follows that the level of significance of a Kolmogorov-Smirnov test with $c = 1.36$ will be 0.05.

Example 10.6.2 Testing Whether a Sample Comes from a Standard Normal Distribution. Suppose that it is desired to test the null hypothesis that a certain random sample of 25 observations was drawn from a standard normal distribution against the alternative that the random sample was drawn from some other continuous distribution. The 25 observed values in the sample, in order from the smallest to the largest, are designated as $y_1, \ldots, y_{25}$ and are listed in Table 10.33. The table also includes the value $F_n(y_i)$ of the sample c.d.f. and the value $\Phi(y_i)$ of the c.d.f. of the standard normal distribution.

By examining the values in Table 10.33, we find that D_n^*, which is the largest difference between $F_n(x)$ and $\Phi(x)$, occurs when we pass from $i = 4$ to $i = 5$, that is, as x increases from the point $x = -0.99$ toward the point $x = -0.42$. The comparison of $F_n(x)$ and $\Phi(x)$ over this interval is illustrated in Fig. 10.4, from which we

Table 10.33 Calculations for Kolmogorov-Smirnov test

i	y_i	$F_n(y_i)$	$\Phi(y_i)$
1	−2.46	0.04	0.0069
2	−2.11	0.08	0.0174
3	−1.23	0.12	0.1093
4	−0.99	0.16	0.1611
5	−0.42	0.20	0.3372
6	−0.39	0.24	0.3483
7	−0.21	0.28	0.4168
8	−0.15	0.32	0.4404
9	−0.10	0.36	0.4602
10	−0.07	0.40	0.4721
11	−0.02	0.44	0.4920
12	0.27	0.48	0.6064
13	0.40	0.52	0.6554
14	0.42	0.56	0.6628
15	0.44	0.60	0.6700
16	0.70	0.64	0.7580
17	0.81	0.68	0.7910
18	0.88	0.72	0.8106
19	1.07	0.76	0.8577
20	1.39	0.80	0.9177
21	1.40	0.84	0.9192
22	1.47	0.88	0.9292
23	1.62	0.92	0.9474
24	1.64	0.96	0.9495
25	1.76	1.00	0.9608

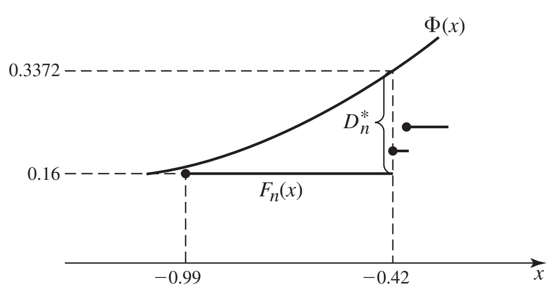


Figure 10.4 The value of D_n^* in Example 10.6.2.

see that $D_n^* = 0.3372 - 0.16 = 0.1772$. Since $n = 25$ in this example, it follows that $n^{1/2}D_n^* = 0.886$. From Table 10.32, we find that $H(0.886) = 0.6$. Hence, the tail area corresponding to the observed value of $n^{1/2}D_n^*$ is 0.4, and we would not reject the null hypothesis at levels α_0 smaller than 0.4. ◀

It is important to emphasize again that when the sample size n is large, even a small value of the tail area corresponding to the observed value of $n^{1/2}D_n^*$ would not necessarily indicate that the true c.d.f. $F(x)$ was much different from the hypothesized c.d.f. $\Phi(x)$. When n itself is large, even a small difference between the c.d.f. $F(x)$ and the c.d.f. $\Phi(x)$ would be sufficient to generate a large value of $n^{1/2}D_n^*$. Therefore, before a statistician rejects the null hypothesis, he should make certain that there is a plausible alternative c.d.f. with which the sample $F_n(x)$ provides closer agreement.

The Kolmogorov-Smirnov Test for Two Samples

Example 10.6.3 Calcium Supplements and Blood Pressure. Exercise 10 in Sec. 9.6 contains data from a study of the effect of a calcium supplement on blood pressure. A group of $m = 10$ men received a calcium supplement, and another group of $n = 11$ men received a placebo. At the end of the study, the differences were calculated between each man's blood pressures at the start and at the end of a 12-week period. Suppose that we are not willing to assume that the distributions of the measured differences are normal distributions. Can we still construct a procedure for testing the null hypothesis that the distributions of differences in the treatment and placebo groups are the same versus the alternative hypothesis that the distributions are different? ◀

Consider a problem in which a random sample of m observations $X_1, \dots, X_m$ is taken from a distribution for which the c.d.f. $F(x)$ is unknown, and an independent random sample of n observations $Y_1, \dots, Y_n$ is taken from another distribution for which the c.d.f. $G(x)$ is also unknown. We shall assume that both $F(x)$ and $G(x)$ are continuous functions and that it is desired to test the hypothesis that these functions are identical, without specifying their common form. Thus, the following hypotheses are to be tested:

$$\begin{aligned} &H_0: \quad F(x) = G(x) \quad \text{for } -\infty < x < \infty,\\ &H_1: \quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{10.6.6}$$

We shall let $F_m(x)$ denote the sample c.d.f. calculated from the observed values of $X_1, \dots, X_m$ and let $G_n(x)$ denote the sample c.d.f. calculated from the observed values of $Y_1, \dots, Y_n$. Furthermore, we shall consider the statistic D_{mn}, which is defined as follows:

$$D_{mn} = \sup_{-\infty<x<\infty} |F_m(x) - G_n(x)|. \tag{10.6.7}$$

The value of D_{mn} is illustrated in Fig. 10.5 for a typical example in which $m = 5$ and $n = 3$.

When the null hypothesis H_0 is true and $F(x)$ and $G(x)$ are identical functions, the sample c.d.f.'s $F_m(x)$ and $G_n(x)$ will tend to be close to each other. In fact, when H_0 is true, it follows from the Glivenko-Cantelli lemma that

$$D_{mn} \xrightarrow{p} 0, \quad \text{as both } m \to \infty \text{ and } n \to \infty. \tag{10.6.8}$$

It seems reasonable, therefore, to use a test procedure that specifies rejecting H_0 when D_{mn} is large. The following theorem, whose proof is beyond the scope of this

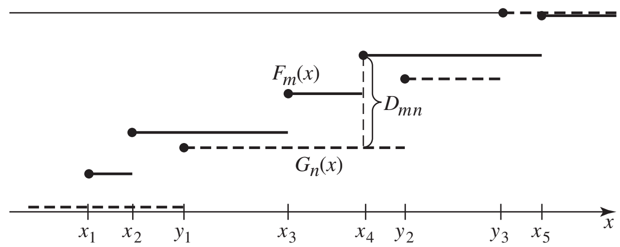


Figure 10.5 A representation of $F_m(x)$, $G_n(x)$, and D_{mn} for $m = 5$ and $n = 3$.

text, gives us the asymptotic distribution of D_{mn}, which we can use to construct an approximate test.

Theorem 10.6.3 Two-Sample Kolmogorov-Smirnov Statistic. For each value of $t > 0$, let $H(t)$ denote the right side of Eq. (10.6.5). If the null hypothesis H_0 in (10.6.6) is true, then

$$\lim_{m\to\infty,\, n\to\infty} \Pr\left[\left(\frac{mn}{m+n}\right)^{1/2} D_{mn} \le t\right] = H(t). \qquad (10.6.9)$$

Values of the function $H(t)$ are given in Table 10.32. The large-sample approximate test of the hypotheses in (10.6.6) makes use of the statistic in (10.6.9).

Definition 10.6.3 Two-Sample Kolmogorov-Smirnov Test. A test procedure that rejects H_0 when

$$\left(\frac{mn}{m+n}\right)^{1/2} D_{mn} \ge c, \qquad (10.6.10)$$

where c is an appropriate constant, is called a *Kolmogorov-Smirnov two-sample test*.

Hence, when the sample sizes m and n are large, the constant c in the relation (10.6.10) can be chosen from Table 10.32 to achieve, at least approximately, any specified level of significance. For example, if m and n are large, and the test is to be carried out at the level of significance 0.05, then it follows from Table 10.32 that we should choose $c = H^{-1}(0.95) = 1.36$.

Example 10.6.4 Calcium Supplements and Blood Pressure. Return to situation described in Example 10.6.3. We are interested in whether or not the changes in blood pressure for men treated with a calcium suppletment have the same distribution as the changes in blood pressure for men treated with a placebo. Figure 10.6 displays the sample c.d.f.'s of the measured changes in the treatment and placebo groups. It is not difficult to see that the maximum difference occurs for $5 \le x < 7$. In fact, $D_{mn} = 0.409$, and the test statistic is $(110/21)^{1/2} \times 0.409 = 0.936$. From Table 10.32, we see that $H(0.936)$ is about 0.654. So we would reject the null hypothesis that the two samples were drawn from the same population at every level $\alpha_0 \ge 0.346$. ◄

Summary

We introduced Kolmogorov-Smirnov tests for testing the null hypotheses that a random sample arose from a particular distribution and that two independent random samples arose from the same distribution. For the one-sample test, we compute D_n, the largest difference between the sample c.d.f. and the null hypothesis c.d.f., and we reject the null hypothesis at level α_0 if $n^{1/2}D_n^* \ge H^{-1}(1-\alpha_0)$, where H is the c.d.f. shown in Table 10.32. For the two-sample test, we compute D_{mn}, the largest

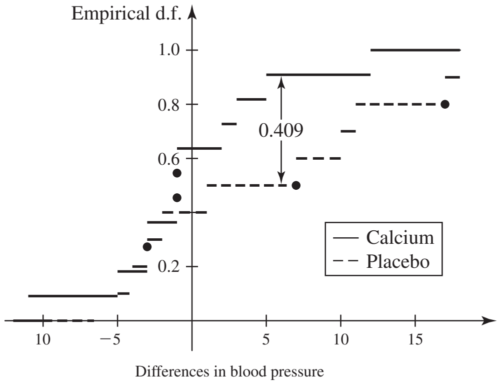


Figure 10.6 The sample c.d.f.'s for two samples in Example 10.6.4.

difference between the two sample c.d.f.'s from the two different samples. We then reject the null hypothesis that the two samples arose from the same distribution at level α_0 if $(mn/(m+n))^{1/2}D_{mn} \geq H^{-1}(1-\alpha_0)$.

Exercises

1. Suppose that the ordered values in a random sample of five observations are $y_1 < y_2 < y_3 < y_4 < y_5$. Let $F_n(x)$ denote the sample c.d.f. constructed from these values, let $F(x)$ be a continuous c.d.f., and let D_n be defined by Eq. (10.6.2). Prove that the minimum possible value of D_n is 0.1, and prove that $D_n = 0.1$ if and only if $F(y_1) = 0.1$, $F(y_2) = 0.3$, $F(y_3) = 0.5$, $F(y_4) = 0.7$, and $F(y_5) = 0.9$.

2. Consider again the conditions of Exercise 1. Prove that $D_n \leq 0.2$ if and only if $F(y_1) \leq 0.2 \leq F(y_2) \leq 0.4 \leq F(y_3) \leq 0.6 \leq F(y_4) \leq 0.8 \leq F(y_5)$.

3. Use the data in Example 10.1.6. In that example, we used a χ^2 goodness-of-fit test to test the null hypothesis that the logarithms of failure times for ball bearings had the normal distribution with mean 3.912 and variance 0.25. Now, use the Kolmogorov-Smirnov test to test that same null hypothesis.

4. Use the Kolmogorov-Smirnov test to test the hypothesis that the 25 values in Table 10.34 form a random sample from the uniform distribution on the interval [0, 1].

Table 10.34 Data for Exercise 4

0.42	0.06	0.88	0.40	0.90
0.38	0.78	0.71	0.57	0.66
0.48	0.35	0.16	0.22	0.08
0.11	0.29	0.79	0.75	0.82
0.30	0.23	0.01	0.41	0.09

5. Use the Kolmogorov-Smirnov test to test the hypothesis that the 25 values given in Exercise 4 form a random sample from the continuous distribution for which the p.d.f. $f(x)$ is as follows:

$$f(x) = \begin{cases} \frac{3}{2} & \text{for } 0 < x \leq \frac{1}{2}, \\ \frac{1}{2} & \text{for } \frac{1}{2} < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

6. Consider again the conditions of Exercise 4 and 5. Suppose that the prior probability is 1/2 that the 25 values given in Table 10.34 were obtained from the uniform distribution on the interval [0, 1], and 1/2 that they were obtained from the distribution for which the p.d.f. is as given in Exercise 5. Find the posterior probability that they were obtained from a uniform distribution.

7. Use the Kolmogorov-Smirnov test to test the hypothesis that the 50 values in Table 10.35 form a random sample from the normal distribution for which the mean is 26 and the variance is 4.

Table 10.35 Data for Exercise 8

25.088	26.615	25.468	27.453	23.845
25.996	26.516	28.240	25.980	30.432
26.560	25.844	26.964	23.382	25.282
24.432	23.593	24.644	26.849	26.801
26.303	23.016	27.378	25.351	23.601
24.317	29.778	29.585	22.147	28.352
29.263	27.924	21.579	25.320	28.129
28.478	23.896	26.020	23.750	24.904
24.078	27.228	27.433	23.341	28.923
24.466	25.153	25.893	26.796	24.743

8. Use the Kolmogorov-Smirnov test to test the hypothesis that the 50 values given in Table 10.35 form a random sample from the normal distribution for which the mean is 24 and the variance is 4.

9. Suppose that 25 observations are selected at random from a distribution for which the c.d.f. $F(x)$ is unknown, and that the values given in Table 10.36 are obtained. Suppose also that 20 observations are selected at random from another distribution for which the c.d.f. $G(x)$ is unknown, and the values given in Table 10.37 are obtained. Use the Kolmogorov-Smirnov test to test the hypothesis that $F(x)$ and $G(x)$ are identical functions.

Table 10.36 First sample for Exercise 9

0.61	0.29	0.06	0.59	−1.73
−0.74	0.51	−0.56	−0.39	1.64
0.05	−0.06	0.64	−0.82	0.31
1.77	1.09	−1.28	2.36	1.31
1.05	−0.32	−0.40	1.06	−2.47

Table 10.37 Second sample for Exercise 9

2.20	1.66	1.38	0.20
0.36	0.00	0.96	1.56
0.44	1.50	−0.30	0.66
2.31	3.29	−0.27	−0.37
0.38	0.70	0.52	−0.71

10. Consider again the conditions of Exercise 9. Let X denote a random variable for which the c.d.f. is $F(x)$, and let Y denote a random variable for which the c.d.f. is $G(x)$. Use the Kolmogorov-Smirnov test to test the hypothesis that the random variables $X + 2$ and Y have the same distribution.

11. Consider again the conditions of Exercises 9 and 10. Use the Kolmogorov-Smirnov test to test the hypothesis that the random variables X and $3Y$ have the same distribution.

12. In Example 9.6.3, we compared two samples of aluminum oxide measurements taken from Roman-era pottery that was found in two different locations in Britain. The $m = 14$ measurements taken from the Llanederyn region are

10.1, 10.9, 11.1, 11.5, 11.6, 12.4, 12.5, 12.7,
13.1, 13.4, 13.8, 13.8, 14.4, 14.6.

The $n = 5$ measurements from Ashley Rails are

14.8, 16.7, 17.7, 18.3, 19.1.

Use the Kolmogorov-Smirnov two-sample test to test the null hypothesis that the two distributions from which these samples are drawn are the same.

13. Suppose that $X_1, \ldots, X_n$ form a random sample with unknown c.d.f. F. Prove the claim made after Eq. (10.6.4) that the distribution of the statistic D_n^*, given that the null hypothesis in (10.6.3) is true, is the same for all continuous F^*. *Hint:* Let $Z_i = F^*(X_i)$ for $i = 1, \ldots, n$, and consider testing the null hypothesis that $Z_1, \ldots, Z_n$ have the uniform distribution on the interval [0, 1]. Show that the statistic D_n^* for this modified problem is identical to the original D_n^*.

14. Perform the Kolmogorov-Smirnov test of the null hypothesis in Example 10.6.1. Report the result of the test by giving the p-value. The sample data appear in Example 10.1.6.

⋆ 10.7 Robust Estimation

In many statistical problems, we might not feel comfortable assuming that the distribution of our data $\boldsymbol{X}$ *is a member of a single parametric family. Suppose that we consider using an estimator* $T = r(\boldsymbol{X})$ *of some parameter* θ. *It might be that* T *has good properties if* $\boldsymbol{X}$ *is a random sample from, say, a normal distribution. On the other hand, we might be concerned about how* T *would behave if* $\boldsymbol{X}$ *were actually a sample from a different distribution. In this section, we introduce a new class of distributions and several new statistics. We then compare the behaviors of these statistics (and some old ones) when the data arise from one of the new distributions (and from some old ones). An estimator is called robust if it performs well, compared to other estimators, regardless of the distribution that gives rise to the data.*

Estimating the Median

Example 10.7.1

Rain from Seeded Clouds. In Fig. 8.3, we presented the histogram of log-rainfalls from 26 seeded clouds, which is slightly asymmetric. A scientist might be uncomfortable treating the log-rainfalls as normal random variables. Nevertheless, one may still wish to estimate the median or some other feature of the distribution of log-rainfalls. One might wish to use a method of estimation that does not rely for its justification on the assumption that the data form a random sample from a normal distribution. ◀

Suppose that the random variables $X_1, \dots, X_n$ form a random sample from a continuous distribution for which the p.d.f. $f(x)$ is unknown, but may be assumed to be a symmetric function with respect to some unknown point θ $(-\infty < \theta < \infty)$. Because of this symmetry, the point θ will be a median of the unknown distribution. We shall estimate the value of θ from the observations $X_1, \dots, X_n$.

If we know that the observations actually come from a normal distribution, then the sample mean $\overline{X}_n$ will be the M.L.E. of θ. Without any strong prior information indicating that the value of θ might be quite different from the observed value of $\overline{X}_n$, we may assume that $\overline{X}_n$ will be a reasonable estimator of θ. Suppose, however, that the observations might come from a distribution for which the p.d.f. $f(x)$ has much thicker tails than the p.d.f. of a normal distribution; that is, suppose that as $x \to \infty$ or $x \to -\infty$, the p.d.f. $f(x)$ might come down to 0 much more slowly than does the p.d.f. of a normal distribution. In this case, the sample mean $\overline{X}_n$ may be a poor estimator of θ because its M.S.E. may be much larger than that of some other possible estimator.

Example 10.7.2

Shifted Cauchy Sample. If the underlying distribution is the Cauchy distribution centered at an unknown point θ, as defined in Example 7.6.5, then the M.S.E. of $\overline{X}_n$ will be infinite. In this case, the M.L.E. of θ will have a finite M.S.E. and will be a much better estimator than $\overline{X}_n$. In fact, for a large value of n, the M.S.E. of the M.L.E. is approximately $2/n$, no matter what the true value of θ is. However, as pointed out in Example 7.6.5, this estimator is very complicated and must be determined by a numerical calculation for each given set of observations. A relatively simple and reasonable estimator for this problem is the *sample median*, which was defined in Example 7.9.3. It can be shown that the M.S.E. of the sample median for a large value of n is approximately $2.47/n$ when the data have the Cauchy distribution. ◀

It follows from Example 10.7.2 and the preceding discussion that if we could assume that the underlying distribution is normal or nearly normal, then we might use the sample mean as an estimator of θ. On the other hand, if we believe that the underlying distribution is Cauchy or nearly Cauchy, then we might use the sample median. However, we typically do not know whether the underlying distribution is nearly normal, is nearly Cauchy, or does not correspond closely to either of these types of distributions. For this reason, we should try to find an estimator of θ that will have a small M.S.E. for several different possible types of distributions. An estimator that performs well for several different types of distributions, even though it may not be the best available estimator for any particular type of distribution, is called a *robust estimator*. In this section, we shall define a class of distributions called *contaminated normals* that we shall use for assessing the performance of various estimators. We shall also introduce special types of robust estimators known as *trimmed means* and *M-estimators*. The term *robust* was introduced by G. E. P. Box in 1953, and the term *trimmed mean* was introduced by J. W. Tukey in 1962. However, the first mathematical treatment of trimmed means was given by P. Daniell in 1920. M-estimators were introduced by Huber (1964).

Contaminated Normal Distributions

One reason that experimenters might be hesitant to behave as if their data were sampled from a normal distribution is the possibility that random errors might occur in the data. Once in a while, a data value is recorded incorrectly or is collected under circumstances that are different from those under study. The one observation (or possibly a few) will have a distribution that might be much different from that of the majority of the observations. For example, suppose that the bulk of the data in which we are interested comprise a sample from the normal distribution with unknown mean μ and variance σ^2. But suppose that, for each observation, there is a small probability ϵ that the observation actually comes from a different distribution with p.d.f. g. That is, the p.d.f. of our observable data is actually

$$f(x) = (1-\epsilon)(2\pi\sigma^2)^{-1/2}\exp\left(-\frac{1}{2\sigma^2}[x-\mu]^2\right) + \epsilon g(x). \tag{10.7.1}$$

Definition 10.7.1 Contaminated Normal Distributions. A distribution whose p.d.f. has the form of Eq. (10.7.1) is called a *contaminated normal*, and the distribution with p.d.f. g is called the *contaminating distribution*.

If the contaminating distribution in Eq. (10.7.1) has a high variance or has a mean very different from μ, there is a good chance that the observations we obtain from the contaminating distribution will be far away from the other observations. In order for an estimator to perform well for a large class of contaminated normal distributions, the estimator will have to be somewhat insensitive to one (or a few) observation(s) not close to the bulk of the data. Obviously, if $\epsilon \geq 1/2$, it becomes difficult to tell which distribution is contaminating which. So we shall assume that $\epsilon < 1/2$. A simple example of a contaminated normal distribution is one in which g is the p.d.f. of a normal distribution with mean μ and variance $100\sigma^2$. In this case, Eq. (10.7.1) becomes

$$\begin{aligned} f(x) = (1-\epsilon)(2\pi\sigma^2)^{-1/2}&\exp\left(-\frac{1}{2\sigma^2}[x-\mu]^2\right) \\ + \epsilon(200\pi\sigma^2)^{-1/2}&\exp\left(-\frac{1}{200\sigma^2}[x-\mu]^2\right). \end{aligned} \tag{10.7.2}$$

Figure 10.7 shows a standard normal p.d.f. together with the p.d.f. of a contaminated normal of the form of Eq. (10.7.2) with $\mu = 0$, $\sigma^2 = 1$, and $\epsilon = 0.05$. The two

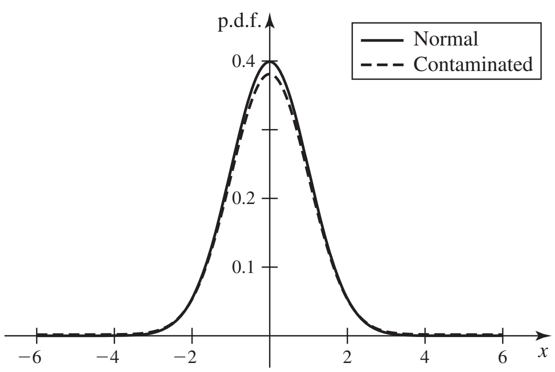


Figure 10.7 p.d.f.'s of standard normal distribution and $\epsilon = 0.05$ contaminated normal with mean 0 and variance 100.

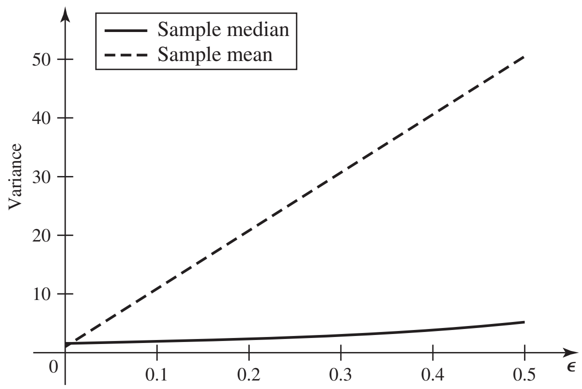


Figure 10.8 Sample size times variances of sample median and sample mean for a random sample from a contaminated normal distribution with the p.d.f. in Eq. (10.7.2) with $\sigma = 1$ as a function of the amount of contamination ϵ. The line for the median uses the asymptotic result Eq. (10.7.3).

p.d.f.'s are quite similar, but we shall see shortly how much effect the contamination can have on the problem of estimation.

Two important properties of the distribution of an estimator of the median are its mean and its variance. In the situation in which the data have the p.d.f. (10.7.2), both the sample mean and the sample median have mean μ. Next, we shall compare the variances of these two estimators when the data are a random sample with the p.d.f. (10.7.2). The variance of the average of a sample of size n is $(1 + 99\epsilon)\sigma^2/n$. (You can prove this in Exercise 7.) The variance of the sample median is a bit more difficult to compute. However, using the large-sample properties that will be introduced on page 676, we can see that the variance is approximately

$$\frac{1}{4nf^2(\mu)} = \frac{\sigma^2}{n}\frac{50\pi}{(10 - 9\epsilon)^2}. \tag{10.7.3}$$

Figure 10.8 shows a comparison of $(50\pi)/(10 - 9\epsilon)^2$ and $(1 + 99\epsilon)$ for $0 \le \epsilon \le 0.5$. Notice that the variance of the sample median is only slightly larger than the variance of the sample mean for $\epsilon < 0.0058$, and it is substantially smaller for ϵ in the range of 0.01 to 0.5. For example, if $\epsilon = 0.05$ (as in Fig. 10.7), the variance of the sample median is only about 29 percent of the variance of the sample mean.

Trimmed Means

Suppose that $X_1, \dots, X_n$ form a random sample from an unknown continuous distribution for which the p.d.f. $f(x)$ is assumed to be symmetric with respect to an unknown point θ. For this discussion, we shall let $Y_1 < Y_2 < \cdots < Y_n$ denote the order statistics of the sample. The sample mean $\overline{X}_n$ is simply the average of these n order statistics. However, if we suspect that the p.d.f. $f(x)$ might have thicker tails than a normal distribution has, then we may wish to estimate θ by using a weighted average of the order statistics, which assigns less weight to the extreme observations such as Y_1, Y_2, Y_{n-1}, and Y_n, and assigns more weight to the middle observations. The sample median is a special example of a weighted average. When n is odd, it assigns zero weight to every observation except the middle one. When n is even, it assigns the weight 1/2 to each of the two middle observations and zero weight to all other observations.

The following class of estimators also consists of weighted averages of the order statistics.

Definition 10.7.2 Trimmed Means. For each positive integer k such that $k < n/2$, ignore the k smallest observations $Y_1, \dots, Y_k$ and the k largest observations $Y_n, Y_{n-1}, \dots, Y_{n-k+1}$ in the sample. The average of the remaining $n - 2k$ intermediate observations is called the kth *level trimmed mean*.

Clearly, the kth level trimmed mean can be represented as a weighted average of the order statistics having the form

$$\frac{1}{n-2k} \sum_{i=k+1}^{n-k} Y_i. \tag{10.7.4}$$

The sample median is an example of a trimmed mean. When n is odd, the sample median is the $[(n-1)/2]$th level trimmed mean. When n is even, it is the $[(n-2)/2]$th level trimmed mean. In either case, the sample median is the kth level trimmed mean, where $k = \lfloor (n-1)/2 \rfloor$ is the largest integer less than or equal to $(n-1)/2$.

Robust Estimation of Scale

In addition to the median of a distribution, there are other parameters that might be worth estimating even when we are not willing to model our data as arising from a particular parametric family. For example, scale parameters might be valuable for giving an idea of how spread out a distribution is. The standard deviation, if it exists, is one such measure. The general class of scale parameters is defined here.

Definition 10.7.3 Scale Parameters. An arbitrary parameter σ is a *scale parameter* for the distribution of X if, for all $a > 0$ and all real b, the corresponding parameter for the distribution of $aX + b$ is $a\sigma$.

Although the standard deviation is a scale parameter, there are many distributions (such as the Cauchy) for which the standard deviation does not exist. There are alternative measures of spread to the standard deviation that exist and are finite for all distributions.

One scale parameter that exists for every distribution is the interquartile range (IQR) as defined in Definition 4.3.2 on page 233. For example, if F is the normal distribution with mean μ and variance σ^2, then the IQR is $2\Phi^{-1}(0.75)\sigma = 1.349\sigma$ (see Exercise 15). The IQR of the Cauchy distribution is 2 (see Example 4.3.9). It is not difficult to show (see Exercise 11) that if the IQR of X is σ and if $a > 0$, then $aX + b$ has IQR equal to $a\sigma$. An estimator of the IQR is the sample IQR, the difference between the 0.75 and 0.25 sample quantiles. (Sample quantiles are just quantiles of the sample c.d.f.)

Another scale parameter that exists for every random variable X is the *median absolute deviation*

Definition 10.7.4 Median Absolute Deviation. The *median absolute deviation* of a random variable X is the median of the distribution of $|X - m|$, where m is the median of X.

If the distribution of X is symmetric around its median, then the median absolute deviation is one-half of the IQR. For asymmetric distributions, the median absolute deviation is the half-length of the symmetric interval around the median that contains 50 percent of the distribution, while the IQR is the length of the interval around the median that contains half of the distribution below the median and half of the distribution above the median. For example, if X has the χ^2 distribution with five

degrees of freedom, the IQR is 3.95, while the median absolute deviation is 1.895, a little less than one-half of the IQR. An estimator of the median absolute deviation is the sample median absolute deviation. The sample median absolute deviation is the sample median of the values $|X_i - M_n|$, where M_n is the sample median of $X_1, \ldots, X_n$.

Two other scale parameters that are useful are the IQR divided by 1.349 and the median absolute deviation divided by 0.6745. These parameters were chosen to have the property that if the data come a normal distribution, then these parameters equal the standard deviation (see Exercise 15). Typical estimators of these parameters are the sample IQR divided by 1.349 and the sample median absolute deviation divided by 0.6745.

M-Estimators of the Median

The sample mean is heavily influenced by one extreme observation. For example, if one observation x in a sample of size n is replaced by $x + \Delta$, the sample mean changes by Δ/n. If Δ is large, this will be a big change. The sample median, on the other hand, is influenced very little, or not at all, by a change in one observation. However, the sample median is inefficient in that it makes use of very few of the observed values. Trimmed means are one attempt to compromise between the sample median and the sample mean by forming estimators that make use of more than just the one or two observations in the middle of the sample while maintaining insensitivity to extreme observations. There are other estimators that also attempt to effect this same type of compromise. These other estimators are M.L.E.'s of θ under different assumptions about the p.d.f. of the observations.

The sample mean is the M.L.E. of θ if we assume that $X_1, \ldots, X_n$ form a random sample from a normal distribution with mean (and median) θ and arbitrary variance. The sample median is also an M.L.E. It is the M.L.E. of θ if we assume that $X_1, \ldots, X_n$ form a random sample from one of the following distributions.

Definition 10.7.5 Laplace Distributions. Let $\sigma > 0$ and θ be real numbers. The distribution whose p.d.f. is

$$f(x|\theta, \sigma) = \frac{1}{2\sigma} e^{-|x-\theta|/\sigma} \tag{10.7.5}$$

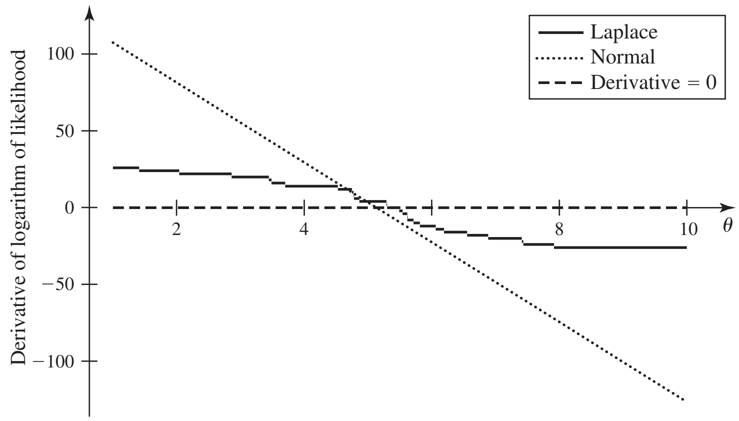
is called the *Laplace distribution with parameters θ and σ.*

See Exercise 9 to prove that the M.L.E. of θ is indeed the sample median when the sample comes from a Laplace distribution.

In order to see why the M.L.E.'s for the Laplace and normal distributions are so different, we can examine the two equations that the M.L.E.'s solve for those two cases. These equations say that the derivatives with respect to θ of the logarithms of the respective likelihoods must equal 0. In both cases, the derivative of the logarithm of the likelihood is the sum of n terms, one for each observation. For the normal case, the term corresponding to an observation x_i is $(x_i - \theta)/\sigma^2$. For the Laplace case, the term corresponding to an observation x_i equals $1/\sigma$ if $\theta < x_i$, and it equals $-1/\sigma$ if $\theta > x_i$. The derivative does not exist at $\theta = x_i$. We illustrate these two derivatives in Fig. 10.9 for the cloud-seeding data introduced in Example 8.3.2. A change of size Δ in a single observation will vertically shift the entire normal distribution line in Fig. 10.9 by $\Delta/[n\sigma^2]$. The same-sized change in the same observation will only affect the Laplace graph in Fig. 10.9 in the vicinity of the changed observation. The actual values of the most extreme observations do not affect where the graph crosses 0.

It would be nice to have a compromise between these two types of behavior without arbitrarily discarding a fixed amount of data. We would like the derivative

Figure 10.9 Derivatives of the logarithms of the Laplace and normal likelihoods (with $\sigma = 1$) using the cloud-seeding data.



of the logarithm of the likelihood to be approximately proportional to $\sum(x_i - \theta)$ for θ near the middle of the data, where the summation is only over the middle observations. This will allow the estimator to make use of more data than just the very middle observation. Also, we would like the derivative to flatten out like the Laplace case for θ near the extremes so that the actual values of the extreme observations do not affect the estimate. A p.d.f. with these properties is the following:

$$g_k(x|\theta, \sigma) = c_k e^{h_k([x-\theta]/\sigma)}, \tag{10.7.6}$$

where σ is a scale parameter,

$$h_k(y) = \begin{cases} -0.5y^2 & \text{if } -k < y < k, \\ 0.5k^2 - k|y| & \text{otherwise,} \end{cases}$$

and c_k is a constant that makes the integral of g equal to 1. The number k must be chosen somehow, usually to reflect some idea of how far from θ we think that extreme observations are likely to be. The derivative of the logarithm of $g_k(x|\theta, \sigma)$ with respect to θ is linear in θ for $|\theta - x| < k\sigma$, but it flattens out like the derivative of the logarithm of the Laplace p.d.f. does when $|\theta - x| > k\sigma$. Now, we see that k can be chosen to reflect how many multiples of σ a data value can be away from θ before we think that it starts to lose importance for estimating θ. Typical choices are $1 \le k \le 2.5$. If we suppose that $X_1, \dots, X_n$ form a random sample from a distribution with p.d.f. $g_k(x|\theta, \sigma)$, the M.L.E. of θ will be a compromise between the sample median and the sample mean.

Definition 10.7.6 *M*-Estimators. The M.L.E. of θ under the assumption that the data have p.d.f. g_k in Eq. (10.7.6) is called an *M-estimator*.

M-estimators were proposed as robust estimators by Huber (1977). The name derives from the fact that they are found by maximizing a function that might not be the likelihood.

The M-estimator found by maximizing $\prod_{i=1}^{n} g_k(x_i|\theta, \sigma)$ cannot be obtained in closed form, but there is a simple iterative algorithm for finding it if we can first estimate σ. Typically, one replaces σ by $\hat{\sigma}$ equal to one of the robust scale estimates described earlier in this section. One popular choice is the sample median absolute deviation divided by 0.6745. Treating $\prod_{i=1}^{n} g_k(x_i|\theta, \hat{\sigma})$ as a function of θ, we can take the derivative of the logarithm and set it equal to 0 to try to find the maximum. The

derivative of the logarithm is $-\sum_{i=1}^{n}\psi_k([x_i-\theta]/\hat{\sigma})/\hat{\sigma}$, where

$$\psi_k(y)=\begin{cases}-k & \text{if } y<-k,\\ y & \text{if } -k\le y\le k,\\ k & \text{if } y>k.\end{cases}$$

Typically, one solves $\sum_{i=1}^{n}\psi_k([x_i-\theta]/\hat{\sigma})=0$ as follows: Rewrite the equation as $\sum_{i=1}^{n}w_i(\theta)(x_i-\theta)=0$, where $w_i(\theta)$ is defined as

$$w_i(\theta)=\begin{cases}\dfrac{\psi_k([x_i-\theta]/\hat{\sigma})}{x_i-\theta} & \text{if } x_i\ne\theta,\\ 1 & \text{if } x_i=\theta.\end{cases}$$

Then $\theta=\sum_{i=1}^{n}w_i(\theta)x_i/\sum_{i=1}^{n}w_i(\theta)$ solves the equation. Clearly, we need to know θ before we can compute $w_i(\theta)$, but we can solve the equation iteratively using these steps:

1. Pick a starting value θ_0 such as the sample median and set $j=0$.
2. Let
$$\theta_{j+1}=\frac{\sum_{i=1}^{n}w_i(\theta_j)x_i}{\sum_{i=1}^{n}w_i(\theta_j)}.$$
3. Increment j to $j+1$, and return to step 2.

This procedure will typically converge in a small number of iterations to the M-estimate $\tilde{\theta}$.

The iterative procedure actually makes it clear why $\tilde{\theta}$ is robust and why it is a compromise between the sample mean and the sample median. Note that $\tilde{\theta}$ is a weighted average of the values $x_1,\dots,x_n$. The weight on x_i is proportional to $w_i(\tilde{\theta})$. If $|x_i-\tilde{\theta}|\le k\hat{\sigma}$, then $w_i(\tilde{\theta})=1/\hat{\sigma}$. If $|x_i-\tilde{\theta}|>k\hat{\sigma}$, then $w_i(\tilde{\theta})=k/|x_i-\tilde{\theta}|$, which decreases as x_i becomes more extreme. If $\tilde{\theta}$ is near the middle of the distribution (as we would hope it would be), then the observations near the middle of the distribution get more weight in the estimate, and those far away get less weight.

Note: *M*-Estimators and Symmetric Distributions. At the start of this section, we assumed that the unknown p.d.f. f of the data was symmetric about an unknown value θ, which must be the median of the distribution. The M-estimator described above can be calculated even if we do not assume that the data come from a symmetric distribution. However, the M-estimator will not necessarily estimate the median of the distribution if the distribution is not symmetric. Instead, the M-estimator estimates the number γ such that

$$E\left[\psi_k\left(\frac{X_i-\gamma}{\sigma}\right)\right]=0. \tag{10.7.7}$$

If the distribution of X_i is symmetric around θ, then $\gamma=\theta$ will solve Eq. (10.7.7). If the distribution of X_i is not symmetric, then some number other than the median might solve Eq. (10.7.7).

Example 10.7.3

Rain from Seeded Clouds. Using the seeded cloud data again, we shall find the value of the M-estimator with $k=1.5$. We start with the sample median of the log-rainfalls, $\theta_0=5.396$. We also use $\hat{\sigma}$ equal to the median absolute deviation 0.7318 divided by 0.6745, that is, $\hat{\sigma}=1.085$. The six smallest and three largest observations are not within $1.5\hat{\sigma}$ of the sample median. These nine observations each get less weight than the other 17 observations in the calculation of the next iteration. For example,

the smallest observation is 1.411, which gets weight $1.5/|1.411 - 5.396| = 0.3764$, compared to weight 0.9217 for the 17 central observations. The weighted average of the observations is then $\theta_1 = 5.315$. We repeat the weighting and averaging until we get no change. After 10 more iterations, we get $\theta_{11} = 5.283$, which agrees with θ_{10}. ◀

Note: Simultaneous *M*-Estimators Exist for the Median and Scale Parameters. It is possible to estimate the median and a scale parameter simultaneously using a method very similar to that described for M-estimators. That is, instead of just picking a value for $\hat{\sigma}$ in the M-estimator algorithm, we can construct a more complicated algorithm that estimates both the median and a scale parameter. Readers interested in more examples of robust procedures can read Huber (1981) and Hampel et al. (1986).

Comparison of the Estimators

We have mentioned the desirability of using a robust estimator in a situation in which it is suspected that the observations $X_1, \ldots, X_n$ may form a random sample from a distribution for which the tails of the p.d.f. are thicker than the tails of the p.d.f. of a normal distribution. The use of a robust estimator is also desirable when a few of the observations in the sample appear to be unusually large or unusually small. In this situation, a statistician might suspect that most of the observations in the sample came from one normal distribution, whereas the few extreme observations may have come from a different normal distribution with a much larger variance than the first one. (This is the contaminated normal case.) The extreme observations, which are called *outliers*, will substantially affect the value of $\overline{X}_n$ and make it an unreliable estimator of θ. Since the values of these outliers would be given less weight in a robust estimator, the robust estimator will usually be a more reliable estimator than $\overline{X}_n$.

It is acknowledged that a robust estimator will perform better than $\overline{X}_n$ in a situation of the type just described. However, if $X_1, \ldots, X_n$ actually do form a random sample from a normal distribution, then $\overline{X}_n$ will perform better than a robust estimator. Since we are typically not certain which situation obtains in a particular problem, it is important to know how much larger the M.S.E. of a robust estimator will be than the M.S.E. of $\overline{X}_n$ when the actual distribution is normal. In other words, it is important to know how much is lost if we use a robust estimator when the actual distribution is normal. We shall now consider this question.

When $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean θ and variance σ^2, the probability distribution of $\overline{X}_n$ and the probability distribution of each of the robust estimators described in this chapter will be symmetric with respect to the value θ. Therefore, the mean of each of these estimators will be θ, the M.S.E. of each estimator will be equal to its variance, and this M.S.E. will have a certain constant value for each estimator regardless of the true value of θ. The values of several of these M.S.E.'s for a normal distribution when the sample size n is 10 or 20 are presented in Table 10.38. The values in Table 10.38 are from Andrews et al. (1972). They were computed using simulation methods that will be introduced in Chapter 12. It should be noted that when $n = 10$, the trimmed mean for $k = 4$ and the sample median are the same estimator.

It can be seen from Table 10.38 that when the underlying distribution is actually a normal distribution, the M.S.E.'s of the M-estimator and the trimmed means are not much larger than the M.S.E. of $\overline{X}_n$. In fact, when $n = 20$, the M.S.E. of the second-level trimmed mean ($k = 2$), in which four of the 20 observed values in the sample

Table 10.38 Comparison of M.S.E.'s for sample mean and several robust estimators. The data have a normal distribution with variance σ^2. The M.S.E. is the tabulated value times σ^2/n. The M-estimator uses $k = 1.5$ and $\hat{\sigma}$ equal to the sample median absolute deviation divided by 0.6745.

Estimator	$n = 10$	$n = 20$
Sample mean $\overline{X}_n$	1.00	1.00
Trimmed mean for $k = 1$	1.05	1.02
Trimmed mean for $k = 2$	1.12	1.06
Trimmed mean for $k = 3$	1.21	1.10
Trimmed mean for $k = 4$	1.37	1.14
Sample median	1.37	1.50
M-estimator	1.05	1.05

Table 10.39 Comparison of M.S.E.'s for sample mean and several robust estimators. The data have a Cauchy distribution. The M.S.E. is the tabulated value divided by n. The M-estimator uses $k = 1.5$ and $\hat{\sigma}$ equal to the sample median absolute deviation divided by 0.6745.

Estimator	$n = 10$	$n = 20$
Sample mean $\overline{X}_n$	∞	∞
Trimmed mean for $k = 1$	27.22	23.98
Trimmed mean for $k = 2$	8.57	7.32
Trimmed mean for $k = 3$	3.86	4.57
Trimmed mean for $k = 4$	3.66	3.58
Sample median	3.66	2.88
M-estimator	6.00	4.50

are omitted, is only 1.06 times as large as the M.S.E. of $\overline{X}_n$. Even the M.S.E. of the sample median is only 1.5 times that of $\overline{X}_n$. These values illustrate the price of using a robust estimator when one is not needed.

We shall now consider the improvement in the M.S.E. that can be achieved by using a robust estimator when the underlying distribution is not normal. If $X_1, \ldots, X_n$ form a random sample of size n from a Cauchy distribution, then the M.S.E. of $\overline{X}_n$ is infinite. The M.S.E.'s of robust estimators for a Cauchy distribution when the sample size n is 10 or 20 are given in Table 10.39. The values in Table 10.39 are from Andrews et al. (1972).

Finally, the M.S.E.'s for two contaminated normal distributions are illustrated in Table 10.40. The two distributions have p.d.f.'s as in Eq. (10.7.2) with $\epsilon = 0.05$ and

Table 10.40 Comparison of M.S.E.'s for sample mean and several robust estimators. The data consist of $n = 20$ observations from a contaminated normal distribution with p.d.f. (10.7.2) using $\epsilon = 0.05$ and $\epsilon = 0.10$. The M.S.E. is the tabulated value divided by n. The M-estimator uses $k = 1.5$ and $\hat{\sigma}$ equal to the sample median absolute deviation divided by 0.6745.

Estimator	$\epsilon = 0.05$	$\epsilon = 0.1$
Sample mean $\overline{X}_n$	5.95	10.90
Trimmed mean for $k = 1$	1.87	3.92
Trimmed mean for $k = 2$	1.32	2.01
Trimmed mean for $k = 3$	1.27	1.57
Trimmed mean for $k = 4$	1.29	1.50
Sample median	1.62	1.81
M-estimator	1.27	1.58

$\epsilon = 0.1$. The values in Table 10.40 were computed using simulation methods described in Chapter 12.

It can be seen from Tables 10.39 and 10.40 that the M.S.E. of a robust estimator can be substantially smaller than that of $\overline{X}_n$. When a trimmed mean or an M-estimator is to be used as an estimator of θ, it is evident that a specific value of k must be chosen. No general rule for choosing k will be best under all conditions. If there is reason to believe that the p.d.f. $f(x)$ is approximately normal, then θ might be estimated by using a trimmed mean, which is obtained by omitting about 10 or 15 percent of the observed values at each end of the ordered sample. Alternatively, an M-estimator with $k = 2$ or 2.5 could be used. If the p.d.f. $f(x)$ might be far from normal or if several of the observations might be outliers, then the sample median might be used to estimate θ, or one could use an M-estimator with $k = 1$ or 1.5.

We could also compare various scale estimators in a similar fashion. Such a comparison is complicated by the fact that there are several choices of scale parameter to estimate, such as standard deviation, IQR, and median absolute deviation. We shall not present such a comparison here.

Large-Sample Properties of Sample Quantiles

Earlier in this section, we made use of the sample median as well as the sample 0.25 and 0.75 quantiles to estimate the median and scale features of a distribution. The distributions of these, and other, sample quantiles are difficult to derive exactly. Approximations are available to the distributions of sample quantiles if the sample sizes are large. It can be shown that if $X_1, \dots, X_n$ form a large random sample from a continuous distribution for which the p.d.f. is $f(x)$ and for which there is a unique p quantile θ_p, then the distribution of the sample p quantile will be approximately a normal distribution. Specifically, it must be assumed that $f(\theta_p) > 0$.

Theorem 10.7.1 Asymptotic Distribution of Sample Quantile. Under the conditions above, let $\tilde{\theta}_{p,n}$ denote the sample p quantile. Then, as $n \to \infty$, the c.d.f. of $n^{1/2}(\tilde{\theta}_{p,n} - \theta_p)$ will converge to the c.d.f. of the normal distribution with mean 0 and variance $p(1-p)/f^2(\theta_p)$.

In other words, when n is large, the distribution of the sample p quantile $\tilde{\theta}_{p,n}$ will be approximately the normal distribution with mean θ_p and variance $p(1-p)/[nf^2(\theta_p)]$.

Also, suppose that $\tilde{\theta}_{q,n}$ denotes the sample q quantile for some $q > p$, and suppose that θ_q is the unique q quantile of the distribution of the data. Then the joint distribution of $(\tilde{\theta}_{p,n}, \tilde{\theta}_{q,n})$ is approximately the bivariate normal distribution with means θ_p and θ_q, variances $p(1-p)/[nf^2(\theta_p)]$ and $q(1-q)/[nf^2(\theta_q)]$, and covariance $p(1-q)/[nf(\theta_p)f(\theta_q)]$. See Schervish (1995, section 7.2) for a rigorous derivation of these results.

Summary

We have introduced a number of estimators of the median and scale parameters that are more robust than the sample average and sample standard deviation. To say that the new estimators are more robust, we mean that they perform well compared to the old estimators, in terms of M.S.E., regardless of which distribution (in some large class) gives rise to the data. The robust estimators of the median include trimmed means, the sample median, and M-estimators obtained by maximizing a function that is similar to a likelihood function. Robust estimators of scale include the sample interquartile range (IQR), the sample median absolute deviation, and multiples of these that are designed to estimate the standard deviation when the data come from a normal distribution.

Exercises

1. Suppose that a sample comprises the 15 observed values in Table 10.41. Calculate the values of **(a)** the sample mean, **(b)** the trimmed means for $k = 1, 2, 3,$ and 4, **(c)** the sample median, and **(d)** the M-estimator with $k = 1.5$ and $\hat{\sigma}$ equal to the sample median absolute deviation divided by 0.6745.

Table 10.41 Data for Exercise 1

23.0	21.5	63.0
22.5	2.1	22.1
22.4	2.2	21.7
21.7	22.2	22.9
21.3	21.8	22.1

2. Suppose that a sample comprises the 14 observed values in Table 10.42. Calculate the values of **(a)** the sample mean, **(b)** the trimmed means for $k = 1, 2, 3,$ and 4, **(c)** the sample median, and **(d)** the M-estimator with $k = 1.5$ and $\hat{\sigma}$ equal to the sample median absolute deviation divided by 0.6745.

Table 10.42 Data for Exercise 2

1.24	0.36	0.23
0.24	1.78	−2.00
−0.11	0.69	0.24
0.10	0.03	0.00
−2.40	0.12	

3. Suppose that a random sample of $n = 100$ observations is taken from the normal distribution with unknown mean θ and known variance 1, and let $\tilde{\theta}_{.5,n}$ denote the sample median. Determine (approximately) the value of $\Pr(|\tilde{\theta}_{.5,n} - \theta| \le 0.1)$.

4. Suppose that a random sample of $n = 100$ observations is taken from the Cauchy distribution centered at an unknown point θ, and let $\tilde{\theta}_{.5,n}$ denote the sample median. Determine (approximately) the value of $\Pr(|\tilde{\theta}_{.5,n} - \theta| \le 0.1)$.

5. Let $f(x)$ denote the p.d.f. of the contaminated normal distribution given in Eq. (10.7.1) with $\epsilon = 1/2, \sigma^2 = 1$, and g being the p.d.f. of a normal distribution with mean μ and variance 4. Suppose that 100 observations are selected at random from a distribution for which the p.d.f. is $f(x)$. Determine the M.S.E. of the sample mean and (approximately) the M.S.E. of the sample median.

6. Use the data in Table 10.6 on page 640. We want an estimate of the median of the logarithms of sulfur dioxide. Find **(a)** the sample mean, **(b)** the trimmed means for $k = 1, 2, 3$, and 4, **(c)** the sample median, and **(d)** the M-estimator with $k = 1.5$ and $\hat{\sigma}$ equal to the sample median absolute deviation divided by 0.6745.

7. Suppose that $X_1, \ldots, X_n$ are i.i.d. with a distribution that has the p.d.f. in Eq. (10.7.2). Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

a. Prove that $E(\overline{X}_n) = \mu$.

b. Prove that $\text{Var}(\overline{X}_n) = (1 + 99\epsilon)\sigma^2/n$.

8. If Fig. 10.8 were extended all the way to $\epsilon = 1$, the variance of the sample median would rise above the variance of the sample mean. Indeed, the ratio of the two variances would be the same at $\epsilon = 1$ as it is at $\epsilon = 0$. Explain why this should be true.

9. Assume that $X_1, \ldots, X_n$ form a random sample from the distribution with p.d.f. given in Eq. (10.7.5). Prove that the M.L.E. of θ is the sample median. (*Hint:* Let X have c.d.f. equal to the sample c.d.f. of $X_1, \ldots, X_n$. Then apply Theorem 4.5.3.)

10. Let $X_1, \ldots, X_n$ be i.i.d. with the p.d.f. in Eq. (10.7.5). Assume that σ is known. Let θ be between two of the observed values $x_1, \ldots, x_n$. Prove that the derivative of the logarithm of the likelihood at θ equals $1/\sigma$ times the difference between the number of observations greater than θ and the number of observations less than θ.

11. Let X be a random variable with a continuous distribution such that the interquartile range (IQR) is σ. Prove that the IQR of $aX + b$ is $a\sigma$ for all $a > 0$ and all b.

12. Let X be a random variable with a continuous distribution such that the median absolute deviation is σ. Prove that the median absolute deviation of $aX + b$ is $a\sigma$ for all $a > 0$ and all b.

13. Find the median absolute deviation of the Cauchy distribution.

14. Let X have the exponential distribution with parameter λ. Prove that the median absolute deviation of X is smaller than one-half of the IQR. (You can do this without actually calculating the median absolute deviation.)

15. Let X have a normal distribution with standard deviation σ.

a. Prove that the IQR is $2\Phi^{-1}(0.75)\sigma$.

b. Prove that the median absolute deviation is $\Phi^{-1}(0.75)\sigma$.

16. Darwin (1876, p. 16) reported the results of an experiment in which he grew 15 pairs of *Zea mays* (corn) plants. Each pair consisted of a self-fertilized and a cross-fertilized plant that were grown in the same pot. The numbers below are the differences between heights (in eighths of an inch) of the two plants in each pair (cross-fertilized minus self-fertilized).

49, −67, 8, 16, 6, 23, 28, 41, 14, 29, 56, 24, 75, 60, −48

Find the **(a)** the sample mean, **(b)** the trimmed means for $k = 1, 2, 3$, and 4, **(c)** the sample median, and **(d)** the M-estimator with $k = 1.5$ and $\hat{\sigma}$ equal to the sample median absolute deviation divided by 0.6745.

17. Let $X_1, \ldots, X_n$ be a large random sample from a distribution with p.d.f. f. Assume that f is symmetric about the median of the distribution. Find the large-sample distribution of the sample IQR.

⋆ 10.8 Sign and Rank Tests

In this section, we describe some popular nonparametric tests for hypotheses about the median of a distribution or about the difference between two distributions.

One-Sample Procedures

Example 10.8.1 Calorie Counts in Hot Dogs. Consider the $n = 20$ calorie counts for beef hot dogs given in Exercise 7 in Sec. 8.5. Suppose that we are interested in testing hypotheses about the median calorie count, but we are not willing to assume that the calorie counts follow a normal distribution or any other familiar distribution. Are there methods

that are appropriate when we are not willing to make assumptions about the form of the distribution? ◀

Suppose that $X_1, \ldots, X_n$ form a random sample from an unknown distribution. In Chapter 9, we considered the case in which the form of the unknown distribution was known, but there were some specific parameters that were still unknown. For example, the distribution might be a normal distribution with unknown mean and/or variance. Now we shall assume only that the distribution is continuous. Since we shall not assume that the distribution of the data has a mean, then we cannot test hypotheses about the mean of the distribution. However, every continuous distribution has a median μ that satisfies $\Pr(X_i \le \mu) = 0.5$. The median is a popular measure of location for general distributions, and we shall now present a test procedure for testing hypotheses of the form

$$\begin{aligned} H_0&: \quad \mu \le \mu_0, \\ H_1&: \quad \mu > \mu_0. \end{aligned} \tag{10.8.1}$$

The test is based on the following simple fact: $\mu \le \mu_0$ if and only if $\Pr(X_i \le \mu_0) \ge 0.5$. For $i = 1, \ldots, n$, let $Y_i = 1$ if $X_i \le \mu_0$, and let $Y_i = 0$ otherwise. Define $p = \Pr(Y_i = 1)$. Then testing whether $\mu \le \mu_0$ is equivalent to testing whether $p \ge 0.5$. Since $X_1, \ldots, X_n$ are independent, then so too are $Y_1, \ldots, Y_n$. This makes $Y_1, \ldots, Y_n$ a random sample from the Bernoulli distribution with parameter p. We already know how to test the null hypothesis that $p \ge 0.5$. (See Example 9.1.9.) We compute $W = Y_1 + \cdots + Y_n$ and reject the null hypothesis if W is too small. To make the test have level of significance α_0, choose c so that

$$\sum_{w=0}^{c} \binom{n}{w} \left(\frac{1}{2}\right)^n \le \alpha_0 < \sum_{w=0}^{c+1} \binom{n}{w} \left(\frac{1}{2}\right)^n.$$

Then the test would reject H_0 if $W \le c$.

The test that we have just described is called the *sign test* because it is based on the number of observations for which $X_i - \mu_0$ is negative. A similar test can be constructed if we wish to test the hypotheses

$$\begin{aligned} H_0&: \quad \mu = \mu_0, \\ H_1&: \quad \mu \ne \mu_0. \end{aligned}$$

Once again, let $p = \Pr(X_i \le \mu_0)$. The null hypothesis H_0 is now equivalent to $p = 0.5$. To perform the test at level of significance α_0, we would choose a number c such that

$$\sum_{w=0}^{c} \binom{n}{w} \left(\frac{1}{2}\right)^n \le \frac{\alpha_0}{2} < \sum_{w=0}^{c+1} \binom{n}{w} \left(\frac{1}{2}\right)^n.$$

We would then reject H_0 if either $W \le c$ or $W \ge n - c$. We use the symmetric rejection region because the binomial distribution with parameters n and $1/2$ is symmetric about $n/2$.

Example 10.8.2 Calorie Counts in Hot Dogs. Consider again the calorie counts for beef hot dogs in Example 10.8.1. Let μ stand for the median of the distribution of calories in beef hot dogs. Suppose that we are interested in testing the hypotheses $H_0 : \mu = 150$ versus $H_1 : \mu \ne 150$. Since 9 of the 20 calorie counts are below 150, we have $W = 9$. The two-sided p-value for this observation is 0.8238, so we would not reject the null hypothesis at level α_0 unless $\alpha_0 \ge 0.8238$. ◀

The power function of the sign test is easy to compute for each value of $p = \Pr(X_i \le \mu_0)$. For example, for the one-sided test of the hypotheses (10.8.1), the power is

$$\Pr(W \le c) = \sum_{w=0}^{c} \binom{n}{w} p^w (1-p)^{n-w}.$$

Comparing Two Distributions

Example 10.8.3 Comparing Copper Ores. Consider again the comparison of copper ores in Example 9.6.5. Suppose that we are not comfortable assuming that the distributions of copper ores are normal distributions. Can we still test hypotheses about whether the distributions are the same or whether they have the same medians? ◄

Next, we shall consider a problem in which a random sample of m observations $X_1, \dots, X_m$ is taken from a continuous distribution for which the c.d.f. $F(x)$ is unknown, and an independent random sample of n observations $Y_1, \dots, Y_n$ is taken from another continuous distribution for which the c.d.f. $G(x)$ is also unknown. We desire to test the hypotheses

$$\begin{aligned} H_0&: \quad F = G\\ H_1&: \quad F \ne G. \end{aligned} \tag{10.8.2}$$

One way to test the hypotheses (10.8.2) is to use the Kolmogorov-Smirnov test for two samples described in Sec. 10.6. Furthermore, if we are willing to assume that the two samples are actually drawn from normal distributions with the same unknown variance, then testing the hypotheses (10.8.2) is the same as testing whether two normal distributions have the same mean. Therefore, under this assumption, we could use a two-sample t test as described in Sec. 9.6.

In this section we shall present another procedure for testing the hypotheses (10.8.2). This procedure, which was introduced separately by F. Wilcoxon and by H. B. Mann and D. R. Whitney in the 1940s, is known as the *Wilcoxon-Mann-Whitney ranks test*.

The Wilcoxon-Mann-Whitney Ranks Test In this procedure, we begin by arranging the $m + n$ observations in the two samples in a single sequence from the smallest value that appears in the two samples to the largest value that appears. Since all the observations come from continuous distributions, it may be assumed that no two of the $m + n$ observations have the same value. Thus, a total ordering of these $m + n$ values can be obtained. Each observation in this total ordering is then assigned a rank from 1 to $m + n$ corresponding to its position in the ordering.

The Wilcoxon-Mann-Whitney ranks test is based on the property that if the null hypothesis H_0 is true and the two samples are actually drawn from the same distribution, then the observations $X_1, \dots, X_m$ will tend to be dispersed throughout the ordering of all $m + n$ observations, rather than be concentrated among the smaller values or among the larger values. In fact, when H_0 is true, the ranks that are assigned to the m observations $X_1, \dots, X_m$ will be the same as if they were a random sample of m ranks drawn at random without replacement from a box containing the $m + n$ ranks $1, 2, \dots, m + n$.

Let S denote the sum of the ranks that are assigned to the m observations $X_1, \dots, X_m$. Since the average of the ranks $1, 2, \dots, m + n$ is $(1/2)(m + n + 1)$, it

Table 10.43 Sorted data for Example 10.8.4

Observed			Observed		
Rank	Value	Sample	Rank	Value	Sample
1	2.120	y	10	2.431	x
2	2.153	y	11	2.556	x
3	2.183	x	12	2.558	y
4	2.213	y	13	2.587	y
5	2.240	y	14	2.629	x
6	2.245	y	15	2.641	x
7	2.266	y	16	2.715	x
8	2.281	y	17	2.805	x
9	2.336	y	18	2.840	x

follows from the discussion just given that when H_0 is true,

$$E(S) = \frac{m(m+n+1)}{2}. \tag{10.8.3}$$

Also, it can be shown that when H_0 is true,

$$\text{Var}(S) = \frac{mn(m+n+1)}{12}. \tag{10.8.4}$$

Furthermore, when the sample sizes m and n are large and H_0 is true, the distribution of S will be approximately the normal distribution for which the mean and the variance are given by Eqs. (10.8.3) and (10.8.4). The Wilcoxon-Mann-Whitney ranks test rejects H_0 if the value of S deviates very far from its mean value given by Eq. (10.8.3). In other words, the test specifies rejecting H_0 if $|S - (1/2)m(m+n+1)| \geq c$, where the constant c is chosen appropriately. In particular, when the approximate normal distribution of S is used, the constant $c = [\text{Var}(S)]^{1/2}\Phi^{-1}(1-\alpha_0/2)$ makes the test have level of significance α_0.

Example 10.8.4 Comparing Copper Ores. Consider again the comparison of copper ores in Example 10.8.3. Suppose that the $m = 8$ measurements in the first sample are

2.183, 2.431, 2.556, 2.629, 2.641, 2.715, 2.805, 2.840,

while the $n = 10$ measurements in the second sample are

2.120, 2.153, 2.213, 2.240, 2.245, 2.266, 2.281, 2.336, 2.558, 2.587.

The 18 values in the two samples are ordered from smallest to largest in Table 10.43. Each observed value in the first sample is identified by the symbol x, and each observed value in the second sample is identified by the symbol y. The sum S of the ranks of the 10 observed values in the first sample is found to be 104.

Suppose that we use the normal distribution approximation. Then if H_0 is true, S has approximately the normal distribution with mean 76 and variance 126.67. The standard deviation of S is therefore $(126.67)^{1/2} = 11.25$. Hence, if H_0 is true, the random variable $Z = (S - 76)/(11.25)$ will have approximately the standard normal distribution. Since $S = 104$ in this example, it follows that $Z = 2.49$. The p-value

corresponding to this value of Z is 0.0128. Hence, the null hypothesis would be rejected at every level of significance $\alpha_0 \geq 0.0128$. ◀

For small values of m and n, the normal approximation to the distribution of S will not be appropriate. Tables of the exact distributions of S for small sample sizes are given in many published collections of statistical tables. Many statistical software packages also calculate the c.d.f. and quantiles of the exact distribution of S.

Note: Tests for Paired Data. Versions of the sign test and ranks test for paired data are developed in Exercises 1 and 15.

Ties

The theory of the Wilcoxon-Mann-Whitney signed ranks test is based on the assumption that all of the observed values of the X_i and Y_j will be distinct. Since the measurements in an actual experiment may be made with only limited precision, however, there may actually be observed values that appear more than once. For example, suppose that a Wilcoxon-Mann-Whitney ranks test is to be performed, and it is found that $X_i = Y_j$ for one or more pairs (i, j). In this case, the ranks test should be carried out twice. In the first test, for each pair with $X_i = Y_j$, it should be assumed that each $X_i < Y_j$. In the second test, assume that $X_i > Y_j$. If the tail areas found from the two tests are roughly equal, then the ties are a relatively unimportant part of the data. If, on the other hand, the tail areas are quite different, then the ties can seriously affect the inferences that are to be made. In this case the data may be inconclusive.

Example 10.8.5

Calcium Supplements and Blood Pressure. Consider the data from Exercise 10 in Sec. 9.6, which we used to illustrate the Kolmogorov-Smirnov test in Example 10.6.4. The observed values −5 and −3 appear in both samples. First, we shall assign the smaller ranks to those values in the group that received the calcium supplement (the X_i's) and then assign the smaller rank to the placebo group (the Y_j's). For example, in the combined sample, the −3 values are the fifth, sixth, and seventh smallest. In the first test, we shall assign rank 5 to the X_i that equals −3 and ranks 6 and 7 to the two Y_j's that equal −3. In the second test, we shall assign rank 7 to the X_i that equals −3 and ranks 5 and 6 to the Y_j's. For the first test, the sum of the X ranks is 123, and in the second test, the sum of the X ranks is 126. In this problem, $m = 10$ and $n = 11$, so the mean and variance of S when the null hypothesis is true are 110 and 201.7, respectively. The two-sided tail areas corresponding to the two assignments are 0.36 and 0.26. Neither of these would lead to rejecting the null hypothesis at level α_0 unless $\alpha_0 \geq 0.26$. ◀

Other reasonable methods for handling ties have been proposed. When two or more values are the same, one simple method is to consider the successive ranks that are to be assigned to these values and then assign the average of these ranks to each of the tied values. When this method is used, the value of $\text{Var}(S)$ must be corrected because of the ties.

Power of the Wilcoxon-Mann-Whitney Ranks Test

The Wilcoxon-Mann-Whitney ranks test rejects the null hypothesis that the two distributions are the same when the sum S of the X ranks is either too large or too small. This would be a sensible thing to do if one thought that the most important

alternatives were those in which the X_i values tended to be larger than the Y_j values or those in which the X_i values tended to be smaller than the Y_j values. However, there are other situations in which $F \neq G$, but S tends to be close to the mean in Eq. (10.8.3). For example, suppose that all $X_1, \ldots, X_m$ have the uniform distribution on the interval [0, 1] and $Y_1, \ldots, Y_n$ have the following p.d.f.:

$$g(y) = \begin{cases} 0.5 & \text{if } -1 < y < 0 \text{ or } 1 < y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Then it is not difficult to show that $E(S)$ is the same as Eq. (10.8.3) and $\text{Var}(S) = m^2 n/4$. In such a case, the power of the test (the probability of rejecting H_0) would not be much larger than the level of significance α_0. Indeed, if one were concerned about alternatives of this sort, one would wish to reject H_0 if the X ranks were too closely clustered regardless of whether they were large or small.

The Wilcoxon-Mann-Whitney ranks test is designed to have high power when F and G have a special relationship to each other, defined next.

Definition 10.8.1 Stochastically Larger. Let X be a random variable with c.d.f. F, and let Y be a random variable with c.d.f. G. Let F^{-1} and G^{-1} denote the respective quantile functions. We say that F is *stochastically larger* than G or, equivalently, that X is *stochastically larger* than Y if $F^{-1}(p) \geq G^{-1}(p)$ for all $0 < p < 1$; that is, every quantile of X is at least as large as the corresponding quantile of Y.

It is easy to see that if X_i is stochastically larger than Y_j, then the ranks of the X_i's in the combined sample will tend to be at least as large as the ranks of the Y_j's. This will make large values of S more likely than small values. Similarly, if Y_j is stochastically larger than X_i, S will tend to be small.

When neither X_i nor Y_j is stochastically larger than the other, it is difficult to make any general claim about the distribution of S. For large sample sizes, a normal approximation still holds for the distribution of S, even when $F \neq G$. However, the mean and variance of S depend on the two c.d.f.'s F and G. For example, using the result in Exercise 11, one can show that

$$E(S) = nm \Pr(X_1 \geq Y_1) + \frac{m(m+1)}{2}. \tag{10.8.5}$$

Using this same approach, one can also show that

$$\begin{aligned} \text{Var}(S) = nm\big[&\Pr(X_1 \geq Y_1) + (1 + m + n) \Pr(X_1 \geq Y_1)^2 \\ &+ (m-1) \Pr(X_1 \geq Y_1, X_1 \geq Y_2) + (n-1) \Pr(X_1 \geq Y_1, X_2 \geq Y_2)\big]. \end{aligned} \tag{10.8.6}$$

In principle, all of these probabilities could be computed for each specific choice of F and G. For particular choices of F and G, one could use simulation methods (see Chapter 12) to approximate the necessary probabilities. After computing or approximating these probabilities, one can then approximate the power of the level α_0 Wilcoxon-Mann-Whitney ranks test as follows: First, recall that the test rejects the null hypothesis that $F = G$ if $S \leq c_1$ or $S \geq c_2$, where

$$c_1 = \frac{m(m+n+1)}{2} - \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)\left[\frac{mn(m+n+1)}{12}\right]^{1/2},$$

$$c_2 = \frac{m(m+n+1)}{2} + \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)\left[\frac{mn(m+n+1)}{12}\right]^{1/2}.$$

Then the power of the test is

$$\Phi\left(\frac{c_1 - E(S)}{\mathrm{Var}(S)^{1/2}}\right) + 1 - \Phi\left(\frac{c_2 - E(S)}{\mathrm{Var}(S)^{1/2}}\right),$$

where $E(S)$ and $\mathrm{Var}(S)$ are given by Eqs. (10.8.5) and (10.8.6), respectively.

Summary

The sign test was introduced as a nonparametric test for hypotheses about the median of an unknown distribution. The Wilcoxon-Mann-Whitney ranks test was developed as another nonparametric test for hypotheses about the equality of two c.d.f.'s. The Wilcoxon-Mann-Whitney ranks test was designed to have large power function when one of the two distributions is stochastically larger than the other.

Exercises

1. Suppose that $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. pairs of random variables with a continuous joint distribution. Let $p = \Pr(X_i \le Y_i)$, and suppose that we want to test the hypotheses

$$\begin{aligned} H_0&: \quad p \le 1/2, \\ H_1&: \quad p > 1/2. \end{aligned} \tag{10.8.7}$$

Describe a version of the sign test to use for testing these hypotheses.

2. Consider again the data in Example 10.8.4. Test the hypotheses (10.8.2) by applying the Kolmogorov-Smirnov test for two samples.

3. Consider again the data in Example 10.8.4. Test the hypotheses (10.8.2) by assuming that the observations are taken from two normal distributions with the same variance, and apply a t test of the type described in Sec. 9.6.

4. In an experiment to compare the effectiveness of two drugs A and B in reducing blood glucose concentrations, drug A was administered to 25 patients, and drug B was administered to 15 patients. The reductions in blood glucose concentrations for the 25 patients who received drug A are given in Table 10.44. The reductions in concentrations for the 15 patients who received drug B are given in Table 10.45. Test the hypothesis that the two drugs are equally effective in reducing blood glucose concentrations by using the Wilcoxon-Mann-Whitney ranks test.

Table 10.44 Data for patients who receive drug A in Exercise 4

0.35	1.12	1.54	0.13	0.77
0.16	1.20	0.40	1.38	0.39
0.58	0.04	0.44	0.75	0.71
1.64	0.49	0.90	0.83	0.28
1.50	1.73	1.15	0.72	0.91

Table 10.45 Data for patients who receive drug B in Exercise 4

1.78	1.25	1.01
1.82	1.95	1.81
0.68	1.48	1.59
0.89	0.86	1.63
1.26	1.07	1.31

5. Consider again the data in Exercise 4. Test the hypothesis that the two drugs are equally effective by applying the Kolmogorov-Smirnov test for two samples.

6. Consider again the data in Exercise 4. Test the hypothesis that the two drugs are equally effective by assuming that the observations are taken from two normal distributions with the same variance and applying a t test of the type described in Sec. 9.6.

7. Suppose that $X_1, \dots, X_m$ form a random sample of m observations from a continuous distribution for which the p.d.f. $f(x)$ is unknown, and that $Y_1, \dots, Y_n$ form an independent random sample of n observations from another continuous distribution for which the p.d.f. $g(x)$ is also unknown. Suppose also that $f(x) = g(x - \theta)$ for $-\infty < x < \infty$, where the value of the parameter θ is unknown $(-\infty < \theta < \infty)$. Let F^{-1} be the quantile function of the X_i's, and let G^{-1} be the quantile function of the Y_j's. Show that $F^{-1}(p) = \theta + G^{-1}(p)$ for all $0 < p < 1$.

8. Consider again the conditions of Exercise 7. Describe how to carry out a one-sided Wilcoxon-Mann-Whitney ranks test of the following hypotheses:

$$\begin{aligned} H_0&: \quad \theta \le 0, \\ H_1&: \quad \theta > 0. \end{aligned}$$

9. Consider again the conditions of Exercise 7. Describe how to carry out a two-sided Wilcoxon-Mann-Whitney ranks test of the following hypotheses for a specified value of θ_0:

$$H_0:\ \theta = \theta_0,$$
$$H_1:\ \theta \neq \theta_0.$$

10. Consider again the conditions of Exercise 9. Describe how to use the Wilcoxon-Mann-Whitney ranks test to determine a confidence interval for θ with confidence coefficient $1 - \alpha_0$. *Hint:* For which values of θ_0 would you accept the null hypothesis $H_0 : \theta = \theta_0$ at level of significance α_0?

11. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be the observations in two samples, and suppose that no two of these observations are equal. Consider the mn pairs

$$\begin{array}{ccc} (X_1, Y_1) & \ldots & (X_1, Y_n), \\ (X_2, Y_1) & \ldots & (X_2, Y_n), \\ & \vdots & \\ (X_m, Y_1) & \ldots & (X_m, Y_n). \end{array}$$

Let U denote the number of these pairs for which the value of the X component is greater than the value of the Y component. Show that

$$U = S - \frac{1}{2}m(m+1),$$

where S is the sum of the ranks assigned to $X_1, \ldots, X_m$, as defined in this section.

12. Let $X_1, \ldots, X_m$ be i.i.d. with c.d.f. F independently of $Y_1, \ldots, Y_n$, which are i.i.d. with c.d.f. G. Let S be as defined in this section. Prove that Eq. (10.8.5) gives the mean of S.

13. Under the conditions of Exercise 12, prove that Eq. (10.8.6) gives the variance of S.

14. Under the conditions of Exercises 12 and 13, suppose further that $F = G$. Prove that Eqs. (10.8.5) and (10.8.6) agree with Eqs. (10.8.3) and (10.8.4), respectively.

15. Consider again the conditions of Exercise 1. This time, let $D_i = X_i - Y_i$. Wilcoxon (1945) developed the following test of the hypotheses (10.8.7). Order the absolute values $|D_1|, \ldots, |D_n|$ from smallest to largest, and assign ranks from 1 to n to the values. Then S_W is set equal to the sum of all the ranks of those $|D_i|$ such that $D_i > 0$. If $p = \Pr(X_i \le Y_i) = 1/2$, then the mean and variance of S_W are

$$E(S_W) = \frac{n(n+1)}{4}, \tag{10.8.8}$$

$$\text{Var}(S_W) = \frac{n(n+1)(2n+1)}{24}. \tag{10.8.9}$$

The test rejects H_0 if $S_W \ge c$, where c is chosen to make the test have level of significance α_0. This test is called the *Wilcoxon signed ranks test*. If n is large, a normal distribution approximation allows us to use $c = E(S_W) + \Phi^{-1}(1 - \alpha_0)\ \text{Var}(S_W)^{1/2}$.

a. Let $W_i = 1$ if $X_i \le Y_i$, and $W_i = 0$ if not. Show that $S_W = \sum_{i=1}^{n} i W_i$.

b. Prove that $E(S_W)$ is as stated in Eq. (10.8.8) under the assumption that $p = 1/2$. *Hint:* You may wish to use Eq. (4.7.13).

c. Prove that $\text{Var}(S_W)$ is as stated in Eq. (10.8.9) under the assumption that $p = 1/2$. *Hint:* You may wish to use Eq. (4.7.14).

16. In an experiment to compare two different materials A and B that might be used for manufacturing the heels of men's dress shoes, 15 men were selected and fitted with a new pair of shoes on which one heel was made of material A and one heel was made of material B. At the beginning of the experiment, each heel was 10 millimeters thick. After the shoes had been worn for one month, the remaining thickness of each heel was measured. The results are given in Table 10.46. Test the null hypothesis that material A is not more durable than material B against the alternative that material A is more durable than material B, by using **(a)** the sign test of Exercise 1, **(b)** the Wilcoxon signed-ranks test of Exercise 15, and **(c)** the paired t test.

Table 10.46 Data for Exercise 16

Pair	Material A	Material B
1	6.6	7.4
2	7.0	5.4
3	8.3	8.8
4	8.2	8.0
5	5.2	6.8
6	9.3	9.1
7	7.9	6.3
8	8.5	7.5
9	7.8	7.0
10	7.5	6.6
11	6.1	4.4
12	8.9	7.7
13	6.1	4.2
14	9.4	9.4
15	9.1	9.1

10.9 Supplementary Exercises

1. Describe how to use the sign test to form a coefficient $1 - \alpha_0$ confidence interval for the median θ of an unknown distribution. Use the data in Exercise 7 in Sec. 8.5 to construct the observed coefficient 0.95 confidence interval. *Hint:* For which values of θ_0 would you fail to reject the null hypothesis $H_0 : \theta = \theta_0$ at level of significance α_0?

2. Suppose that 400 persons are chosen at random from a large population, and that each person in the sample specifies which one of five breakfast cereals she most prefers. For $i = 1, \ldots, 5$, let p_i denote the proportion of the population that prefers cereal i, and let N_i denote the number of persons in the sample who prefer cereal i. It is desired to test the following hypotheses at the level of significance 0.01:

$$H_0: \quad p_1 = p_2 = \cdots = p_5,$$
$$H_1: \quad \text{The hypothesis } H_0 \text{ is not true.}$$

For what values of $\sum_{i=1}^{5} N_i^2$ would H_0 be rejected?

3. Consider a large population of families that have exactly three children, and suppose that it is desired to test the null hypothesis H_0 that the distribution of the number of boys in each family is a binomial distribution with parameters $n = 3$ and $p = 1/2$ against the general alternative H_1 that H_0 is not true. Suppose also that in a random sample of 128 families it is found that 26 families have no boys, 32 families have one boy, 40 families have two boys, and 30 families have three boys. At what levels of significance should H_0 be rejected?

4. Consider again the conditions of Exercise 3, including the observations in the random sample of 128 families, but suppose now that it is desired to test the composite null hypothesis H_0 that the distribution of the number of boys in each family is a binomial distribution for which $n = 3$, and the value of p is not specified against the general alternative H_1 that H_0 is not true. At what levels of significance should H_0 be rejected?

5. In order to study the genetic history of three different large groups of Americans, a random sample of 50 persons is drawn from group 1, a random sample of 100 persons is drawn from group 2, and a random sample of 200 persons is drawn from group 3. The blood type of each person in the samples is classified as A, B, AB, or O, and the results are as given in Table 10.47. Test the hypothesis that the distribution of blood types is the same in all three groups at the level of significance 0.1.

Table 10.47 Data for Exercises 5 and 6

	A	B	AB	O	Total
Group 1	24	6	5	15	50
Group 2	43	24	7	26	100
Group 3	69	47	22	62	200

6. Consider again the conditions of Exercise 5. Explain how to change the numbers in Table 10.47 in such a way that each row total and each column total remains unchanged, but the value of the χ^2 test statistic is increased.

7. Consider a χ^2 test of independence that is to be applied to the elements of a 2×2 contingency table. Show that the quantity $(N_{ij} - \hat{E}_{ij})^2$ has the same value for each of the four cells of the table.

8. Consider again the conditions of Exercise 7. Show that the χ^2 statistic Q can be written in the form

$$Q = \frac{n(N_{11}N_{22} - N_{12}N_{21})^2}{N_{1+}N_{2+}N_{+1}N_{+2}}.$$

9. Suppose that a χ^2 test of independence at the level of significance 0.01 is to be applied to the elements of a 2×2 contingency table containing $4n$ observations, and that the data have the form given in Table 10.48. For what values of a would the null hypothesis be rejected?

Table 10.48 Form of the data for Exercise 9

$n + a$	$n - a$
$n - a$	$n + a$

10. Suppose that a χ^2 test of independence at the level of significance 0.005 is to be applied to the elements of a 2×2 contingency table containing $2n$ observations, and that for some $\alpha \in (0, 1)$ the data have the form given in Table 10.49. For what values of α would the null hypothesis be rejected?

Table 10.49 Form of the data for Exercise 10

αn	$(1 - \alpha)n$
$(1 - \alpha)n$	αn

11. In a study of the health effects of air pollution, it was found that the proportion of the total population of city A that suffered from respiratory diseases was larger than the proportion for city B. Since city A was generally regarded as being less polluted and more healthful than city B, this result was considered surprising. Therefore, separate investigations were made for the younger population (under age 40) and for the older population (age 40 or older). It

was found that the proportion of the younger population suffering from respiratory diseases was smaller for city A than for city B, and also that the proportion of the older population suffering from respiratory diseases was smaller for city A than for city B. Discuss and explain these results.

12. Suppose that an achievement test in mathematics was given to students from two different high schools A and B. When the results of the test were tabulated, it was found that the average score for the freshmen at school A was higher than the average for the freshmen at school B, and that the same relationship existed for the sophomores, the juniors, and the seniors at the two schools. On the other hand, it was found also that the average score of all the students at school A was lower than that of all the students at school B. Discuss and explain these results. Give an example of how this could happen.

13. A random sample of 100 hospital patients suffering from depression received a particular treatment over a period of three months. Prior to the beginning of the treatment, each patient was classified as being at one of five levels of depression, where level 1 represented the most severe level of depression and level 5 represented the mildest level. At the end of the treatment, each patient was again classified according to the same five levels of depression. The results are given in Table 10.50. Discuss the use of this table for determining whether the treatment has been helpful in alleviating depression.

Table 10.50 Data for Exercise 13

Level of depression before treatment	Level of depression after treatment				
	1	2	3	4	5
1	7	3	0	0	0
2	1	27	14	2	0
3	0	0	19	8	2
4	0	1	2	12	0
5	0	0	1	1	0

14. Suppose that a random sample of three observations is drawn from a distribution with the following p.d.f.:

$$f(x) = \begin{cases} \theta x^{\theta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$. Determine the p.d.f. of the sample median.

15. Suppose that a random sample of n observations is drawn from a distribution for which the p.d.f. is as given in Exercise 14. Determine the asymptotic distribution of the sample median.

16. Suppose that a random sample of n observations is drawn from a t distribution with $\alpha > 2$ degrees of freedom. Show that the asymptotic distributions of both the sample mean $\overline{X}_n$ and the sample median $\tilde{X}_n$ are normal, and determine the positive integers α for which the variance of the asymptotic distribution is smaller for $\overline{X}_n$ than for $\tilde{X}_n$.

17. Suppose that $X_1, \dots, X_n$ form a large random sample from a distribution for which the p.d.f. is $h(x|\theta) = \alpha f(x|\theta) + (1-\alpha)g(x|\theta)$. Here $f(x|\theta)$ is the p.d.f. of the normal distribution with unknown mean θ and variance 1, $g(x|\theta)$ is the p.d.f. of the normal distribution with the same unknown mean θ and variance σ^2, and $0 \le \alpha \le 1$. Let $\overline{X}_n$ and $\tilde{X}_n$ denote the sample mean and the sample median, respectively.

a. For $\sigma^2 = 100$, determine the values of α for which the M.S.E. of $\tilde{X}_n$ will be smaller than the M.S.E. of $\overline{X}_n$.

b. For $\alpha = 1/2$, determine the values of σ^2 for which the M.S.E. of $\tilde{X}_n$ will be smaller than the M.S.E. of $\overline{X}_n$.

18. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution with p.d.f. $f(x)$, and let $Y_1 < Y_2 < \cdots < Y_n$ denote the order statistics of the sample. Prove that the joint p.d.f. of $Y_1, \dots, Y_n$ is as follows:

$$g(y_1, \dots, y_n) = \begin{cases} n! f(y_1) \cdots f(y_n) & \text{for } y_1 < y_2 < \cdots < y_n \\ 0 & \text{otherwise.} \end{cases}$$

19. Let $Y_1 < Y_2 < Y_3$ denote the order statistics of a random sample of three observations from the uniform distribution on the interval [0, 1]. Determine the conditional distribution of Y_2 given that $Y_1 = y_1$ and $Y_3 = y_3$ $(0 < y_1 < y_3 < 1)$.

20. Suppose that a random sample of 20 observations is drawn from an unknown continuous distribution, and let $Y_1 < \cdots < Y_{20}$ denote the order statistics of the sample. Also, let θ denote the 0.3 quantile of the distribution, and suppose that it is desired to present a confidence interval for θ that has the form (Y_r, Y_{r+3}). Determine the value of r $(r = 1, 2, \dots, 17)$ for which this interval will have the largest confidence coefficient γ, and determine the value of γ.

21. Suppose that $X_1, \dots, X_m$ form a random sample from a continuous distribution for which the p.d.f. $f(x)$ is unknown; $Y_1, \dots, Y_n$ form an independent random sample from another continuous distribution for which the p.d.f. $g(x)$ also is unknown; and $f(x) = g(x - \theta)$ for $-\infty < x < \infty$, where the value of the parameter θ is unknown $(-\infty < \theta < \infty)$. Suppose that it is desired to carry out a Wilcoxon-Mann-Whitney ranks test of the following hypotheses at a specified level of significance α $(0 < \alpha < 1)$:

$$H_0: \quad \theta = \theta_0,$$
$$H_1: \quad \theta \neq \theta_0.$$

Assume that no two of the observations are equal, and

let U_{θ_0} denote the number of pairs (X_i, Y_j) such that $X_i - Y_j > \theta_0$, where $i = 1, \dots, m$ and $j = 1, \dots, n$. Show that for large values of m and n, the hypothesis H_0 should not be rejected if and only if

$$\frac{mn}{2} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\left[\frac{mn(m+n+1)}{12}\right]^{1/2} < U_{\theta_0} < \frac{mn}{2} + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\left[\frac{mn(m+n+1)}{12}\right]^{1/2},$$

where Φ^{-1} is the quantile function of the standard normal distribution. *Hint:* See Exercise 11 of Sec. 10.8.

22. Consider again the conditions of Exercise 21. Show that a confidence interval for θ with confidence coefficient $1 - \alpha$ can be obtained by the following procedure: Let k be the largest integer less than or equal to

$$\frac{mn}{2} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\left[\frac{mn(m+n+1)}{12}\right]^{1/2}.$$

Also, let A be the kth smallest of the mn differences $X_i - Y_j$, where $i = 1, \dots, m$ and $j = 1, \dots, n$, and let B be the kth largest of these mn differences. Then the interval $A < \theta < B$ is a confidence interval of the required type.

23. The sign test can be extended to a test of hypotheses about an arbitrary quantile of a distribution rather than just the median. Let θ_p be the p quantile of a distribution, and suppose that $X_1, \dots, X_n$ form an i.i.d. sample from this distribution.

a. Let b be an arbitrary number. Explain how to construct a version of the sign test for the hypotheses

$$H_0: \quad \theta_p = b,$$
$$H_1: \quad \theta_p \neq b,$$

at level of significance α_0. (Construct an equal-tailed test if you wish.)

b. Show how to use this version of the sign test to form a coefficient $1 - \alpha_0$ confidence interval for θ_p.

Chapter

11

LINEAR STATISTICAL MODELS

11.1 The Method of Least Squares

When each observation from an experiment is a pair of numbers, it is often important to try to predict one of the numbers from the other. Least squares is a method for constructing a predictor of one of the variables from the other by making use of a sample of observed pairs.

Fitting a Straight Line

Example 11.1.1

Blood Pressure. Suppose that each of 10 patients is treated with the same amount of two different drugs that can affect blood pressure. To be specific, each patient is first treated with a standard drug A, and their change in blood pressure is measured. After the effect of the drug wears off, the patient is treated with an equal amount of a new drug B, and their change in blood pressure is measured again. These changes in blood pressure will be called the *reaction* of the patient to each drug. For $i = 1, \dots, 10$, we shall let x_i denote the reaction, measured in appropriate units, of the ith patient to drug A, and we shall let y_i denote her reaction to drug B. The observed values of the reactions are as given in Table 11.1. The 10 points (x_i, y_i) for $i = 1, \dots, 10$ are plotted in Fig. 11.1. One purpose of the study is to try to predict a patient's reaction to drug B if their reaction to the standard drug A is already known. ◄

In Example 11.1.1, suppose that we are interested in describing the relationship between the reaction y of a patient to drug B and her reaction x to drug A. In order to obtain a simple expression for this relationship, we might wish to fit a straight line to the 10 points plotted in Fig. 11.1. Although these 10 points obviously do not lie exactly on a straight line, we might believe that the deviations from such a line are caused by the fact that the observed change in the blood pressure of each patient is affected not only by the two drugs but also by various other factors. In other words, we might believe that if it were possible to control all of these other factors, the observed points would actually lie on a straight line. We might believe further that if we measured the reactions to the two drugs for a very large number of patients, instead of for just 10 patients, we would then find that the observed points tend to

Table 11.1 Reactions to two drugs

i	x_i	y_i
1	1.9	0.7
2	0.8	−1.0
3	1.1	−0.2
4	0.1	−1.2
5	−0.1	−0.1
6	4.4	3.4
7	4.6	0.0
8	1.6	0.8
9	5.5	3.7
10	3.4	2.0

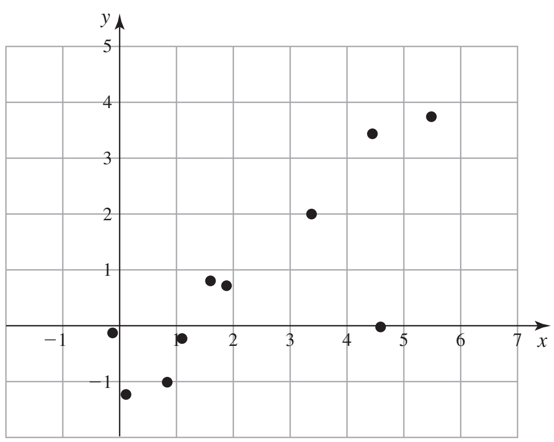


Figure 11.1 A plot of the observed values in Table 11.1.

cluster along a straight line. Perhaps we might also wish to be able to predict the reaction y of a future patient to the new drug B on the basis of her reaction x to the standard drug A. One procedure for making such a prediction would be to fit a straight line to the points in Fig. 11.1, and to use this line for predicting the value of y corresponding to each value of x.

It can be seen from Fig. 11.1 that if we did not have to consider the point (4.6, 0.0), which is obtained from the patient for whom $i = 7$ in Table 11.1, then the other nine points lie roughly along a straight line. One arbitrary line that fits reasonably well to these nine points is sketched in Fig. 11.2. However, if we wish to fit a straight line to all 10 points, it is not clear just how much the line in Fig. 11.2 should be adjusted in order to accommodate the anomalous point. We shall now describe a method for fitting such a line.

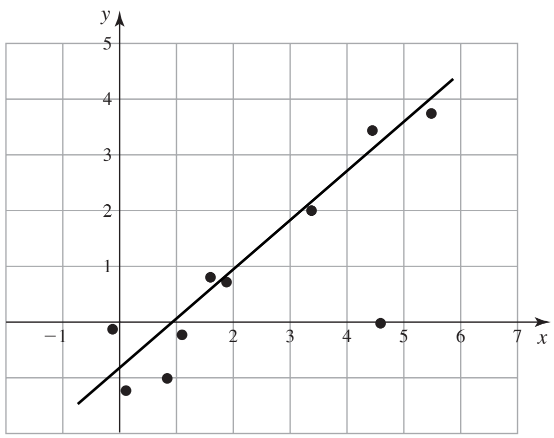


Figure 11.2 A straight line fitted to nine of the points in Table 11.1.

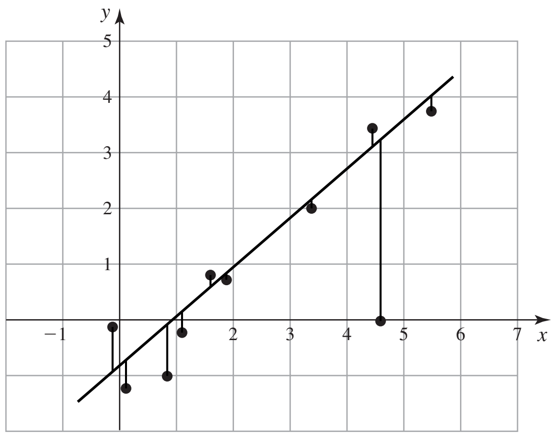


Figure 11.3 Vertical deviations of the plotted points from a straight line.

The Least-Squares Line

Example 11.1.2

Blood Pressure. In Example 11.1.1, suppose that we are interested in fitting a straight line to the points plotted in Fig. 11.1 in order to obtain a simple mathematical relationship for expressing the reaction y of a patient to the new drug B as a function of her reaction x to the standard drug A. In other words, our main objective is to be able to predict closely a patient's reaction y to drug B from her reaction x to drug A. We are interested, therefore, in constructing a straight line such that, for each observed reaction x_i, the corresponding value of y on the straight line will be as close as possible to the actual observed reaction y_i. The vertical deviations of the 10 plotted points from the line drawn in Fig. 11.2 are sketched in Fig. 11.3. ◀

One method of constructing a straight line to fit the observed values is called *the method of least squares*, which chooses the line to minimize the sum of the squares of the vertical deviations of all the points from the line. We shall now study the method of least squares in more detail.

Theorem 11.1.1 Least Squares. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a set of n points. The straight line that minimzes the sum of the squares of the vertical deviations of all the points from the line has the following slope and intercept:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}, \tag{11.1.1}$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Proof Consider an arbitrary straight line $y = \beta_0 + \beta_1 x$, in which the values of the constants β_0 and β_1 are to be determined. When $x = x_i$, the height of this line is $\beta_0 + \beta_1 x_i$. Therefore, the vertical distance between the point (x_i, y_i) and the line is $|y_i - (\beta_0 + \beta_1 x_i)|$. Suppose that the line is to be fitted to n points. The sum of the squares of the vertical distances at the n points is

$$Q = \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2. \tag{11.1.2}$$

We shall minimize Q with respect to β_0 and β_1 by taking the partial derivatives and setting them to 0. We have

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \tag{11.1.3}$$

and

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i. \tag{11.1.4}$$

By setting each of these two partial derivatives equal to 0, we obtain the following pair of equations:

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \qquad \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{11.1.5}$$

The equations (11.1.5) are called the *normal equations* for β_0 and β_1. By considering the second-order derivatives of Q, we can show that the values of β_0 and β_1 that satisfy the normal equations will be the values for which the sum of squares Q in Eq. (11.1.2) is minimized. Solving (11.1.5) yields the values in (11.1.1). ■

Definition 11.1.1 Least-Squares Line. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be as defined in (11.1.1). The line defined by the equation $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is called the *least-squares line*.

For the values given in Table 11.1, $n = 10$, and it is found from Eq. (11.1.1) that $\hat{\beta}_0 = -0.786$ and $\hat{\beta}_1 = 0.685$. Hence, the equation of the least-squares line is $y = -0.786 + 0.685x$. This line is sketched in Fig. 11.4.

Virtually all statistical computer software will compute the least-squares regression line. Even some handheld calculators will do the calculation.

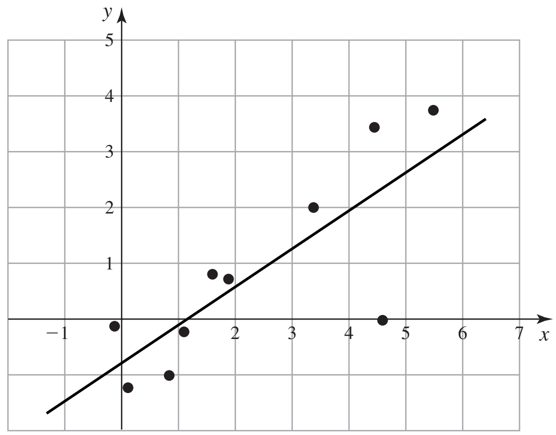


Figure 11.4 The least-squares straight line.

Fitting a Polynomial by the Method of Least Squares

Suppose now that instead of simply fitting a straight line to n plotted points, we wish to fit a polynomial of degree k ($k \geq 2$). Such a polynomial will have the following form:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k. \tag{11.1.6}$$

The method of least squares specifies that the constants $\beta_0, \ldots, \beta_k$ should be chosen so that the sum Q of the squares of the vertical deviations of the points from the curve is a minimum. In other words, these constants should be chosen so as to minimize the following expression for Q:

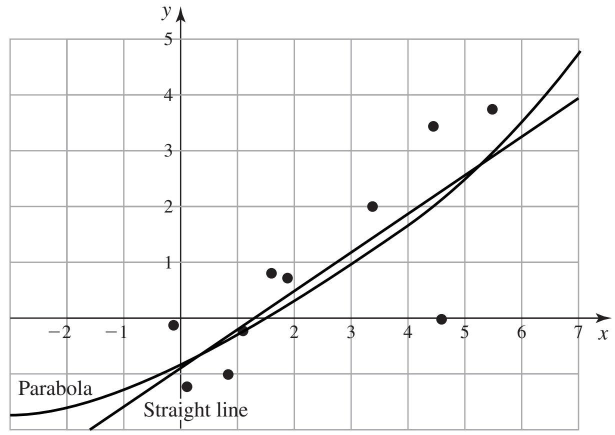
$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i + \cdots + \beta_k x_i^k)]^2. \tag{11.1.7}$$

If we calculate the $k+1$ partial derivatives $\partial Q/\partial\beta_0, \ldots, \partial Q/\partial\beta_k$, and we set each of these derivatives equal to 0, we obtain the following $k+1$ linear equations involving the $k+1$ unknown values $\beta_0, \ldots, \beta_k$:

$$\begin{aligned}
\beta_0 n + \beta_1 \sum_{i=1}^{n} x_i + \cdots + \beta_k \sum_{i=1}^{n} x_i^k &= \sum_{i=1}^{n} y_i, \\
\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 + \cdots + \beta_k \sum_{i=1}^{n} x_i^{k+1} &= \sum_{i=1}^{n} x_i y_i, \\
&\vdots \\
\beta_0 \sum_{i=1}^{n} x_i^k + \beta_1 \sum_{i=1}^{n} x_i^{k+1} + \cdots + \beta_k \sum_{i=1}^{n} x_i^{2k} &= \sum_{i=1}^{n} x_i^k y_i.
\end{aligned} \tag{11.1.8}$$

As before, these equations are called the *normal equations*. If the normal equations have a unique solution, that solution provides the minimum value for Q. A necessary and sufficient condition for a unique solution is that the determinant of the $(k+1) \times (k+1)$ matrix formed by the coefficients of $\beta_0, \ldots, \beta_k$ in Eq. (11.1.8) is not zero. We shall now assume that this is the case. If we denote the solution as $(\hat{\beta}_0, \ldots, \hat{\beta}_k)$, then the least-squares polynomial is $y = \hat{\beta}_0 + \hat{\beta}_1 x + \cdots + \hat{\beta}_k x^k$.

Figure 11.5 The least-squares parabola.



Example 11.1.3

Fitting a Parabola. Suppose that we wish to fit a polynomial of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2$ (which represents a parabola) to the 10 points given in Table 11.1. In this example, it is found that the normal equations 11.1.8 are as follows:

$$\begin{aligned} 10\beta_0 + 23.3\beta_1 + 90.37\beta_2 &= 8.1, \\ 23.3\beta_0 + 90.37\beta_1 + 401.0\beta_2 &= 43.59, \\ 90.37\beta_0 + 401.0\beta_1 + 1892.7\beta_2 &= 204.55. \end{aligned} \quad (11.1.9)$$

The unique values of β_0, β_1, and β_2 that satisfy these three equations are $\hat{\beta}_0 = -0.744$, $\hat{\beta}_1 = 0.616$, and $\hat{\beta}_2 = 0.013$. Hence, the least-squares parabola is

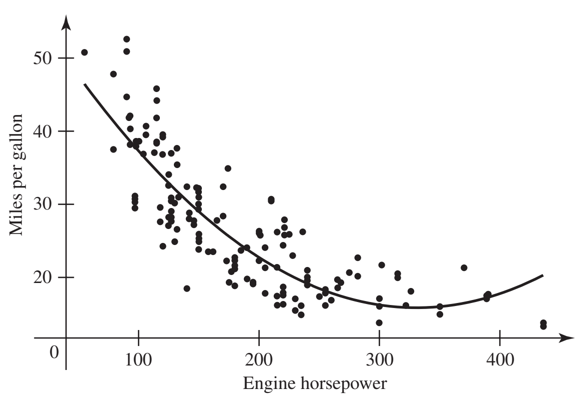
$$y = -0.744 + 0.616x + 0.013x^2. \quad (11.1.10)$$

This curve is sketched in Fig. 11.5 together with the least-squares straight line. Because the coefficient of x^2 in Eq. (11.1.10) is so small, the least-squares parabola and the least-squares straight line are very close together over the range of values included in Fig. 11.5. ◀

Example 11.1.4

Gasoline Mileage. Heavenrich and Hellman (1999) report several variables measured on 173 different cars. Among those variables are gasoline mileage (in miles per gallon) and engine horsepower. A plot of miles per gallon versus horsepower is shown in Fig. 11.6 together with a parabola fit by least squares. Even without the curve

Figure 11.6 Plot of miles per gallon versus engine horsepower for 173 cars in Example 11.1.4. The least-squares parabola is also drawn in the plot.



drawn in Fig. 11.6, it is clear that a straight line would not provide an adequate fit to the relationship between these two variables. Some sort of curved relationship must be fit. The least-squares parabola curves up for the largest values of horsepower, which is somewhat counterintuitive. Indeed, this might be an example in which it would pay to use some prior information to impose a constraint on the fitted curve. Alternatively, we could replace gasoline mileage by a curved function of miles per gallon and use this curved function as the y variable. ◀

Fitting a Linear Function of Several Variables

We shall now consider an extension of the example discussed at the beginning of this section, in which we were interested in representing a patient's reaction to a new drug B as a linear function of her reaction to drug A. Suppose that we wish to represent a patient's reaction to drug B as a linear function involving not only her reaction to drug A but also some other relevant variables. For example, we may wish to represent the patient's reaction y to drug B as a linear function involving her reaction x_1 to drug A, her heart rate x_2, and blood pressure x_3 before she receives any drugs, and other relevant variables $x_4, \dots, x_k$.

Suppose that for each patient i $(i = 1, \dots, n)$ we measure her reaction y_i to drug B, her reaction x_{i1} to drug A, and also her values $x_{i2}, \dots, x_{ik}$ for the other variables. Suppose also that in order to fit these observed values for the n patients, we wish to consider a linear function having the form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{11.1.11}$$

In this case, also, the values of $\beta_0, \dots, \beta_k$ can be determined by the method of least squares. For each given set of observed values $x_{i1}, \dots, x_{ik}$, we again consider the difference between the observed reaction y_i and the value $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ of the linear function given in Eq. (11.1.11). As before, it is required to minimize the sum Q of the squares of these differences. Here,

$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})]^2. \tag{11.1.12}$$

We minimize this the same way that we minimized (11.1.7), namely, by setting the partial derivatives of Q with respect to each β_j equal to 0 for $j = 0, \dots, k$. In this case, the $k + 1$ normal equations have the following form:

$$\begin{aligned}
\beta_0 n + \beta_1 \sum_{i=1}^{n} x_{i1} + \cdots + \beta_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i, \\
\beta_0 \sum_{i=1}^{n} x_{i1} + \beta_1 \sum_{i=1}^{n} x_{i1}^2 + \cdots + \beta_k \sum_{i=1}^{n} x_{i1} x_{ik} &= \sum_{i=1}^{n} x_{i1} y_i, \\
&\vdots \\
\beta_0 \sum_{i=1}^{n} x_{ik} + \beta_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \cdots + \beta_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik} y_i.
\end{aligned} \tag{11.1.13}$$

If the normal equations have a unique solution, we shall denote that solution $(\hat{\beta}_0, \dots, \hat{\beta}_k)$, and the least-squares linear function will then be $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$. As before, a necessary and sufficient condition for a unique solution is that the determinant of the $(k + 1) \times (k + 1)$ matrix formed by the coefficients of $\beta_0, \dots, \beta_k$ in Eq. (11.1.13) is not zero.

Table 11.2 Reactions to two drugs and heart rate

i	x_{i1}	x_{i2}	y_i
1	1.9	66	0.7
2	0.8	62	−1.0
3	1.1	64	−0.2
4	0.1	61	−1.2
5	−0.1	63	−0.1
6	4.4	70	3.4
7	4.6	68	0.0
8	1.6	62	0.8
9	5.5	68	3.7
10	3.4	66	2.0

Example 11.1.5 Fitting a Linear Function of Two Variables. Suppose that we expand Table 11.1 to include the values given in the third column in Table 11.2. Here, for each patient i $(i = 1, \dots, 10)$, x_{i1} denotes her reaction to the standard drug A, x_{i2} denotes her heart rate, and y_i denotes her reaction to the new drug B. Suppose also that we wish to fit a linear function to these values having the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

In this example, it is found that the normal equations (11.1.13) are

$$\begin{aligned} 10\beta_0 + 23.3\beta_1 + 650\beta_2 &= 8.1, \\ 23.3\beta_0 + 90.37\beta_1 + 1563.6\beta_2 &= 43.59, \\ 650\beta_0 + 1563.6\beta_1 + 42,334\beta_2 &= 563.1. \end{aligned} \tag{11.1.14}$$

The unique values of β_0, β_1, and β_2 that satisfy these three equations are $\hat{\beta}_0 = -11.4527$, $\hat{\beta}_1 = 0.4503$, and $\hat{\beta}_2 = 0.1725$. Hence, the least-squares linear function is

$$y = -11.4527 + 0.4503x_1 + 0.1725x_2. \tag{11.1.15}$$

◀

It should be noted that the problem of fitting a polynomial of degree k involving only one variable, as specified by Eq. (11.1.6), can be regarded as a special case of the problem of fitting a linear function involving several variables, as specified by Eq. (11.1.11). To make Eq. (11.1.11) applicable to the problem of fitting a polynomial having the form given in Eq. (11.1.6), we define the k variables $x_1, \dots, x_k$ simply as $x_1 = x$, $x_2 = x^2, \dots, x_k = x^k$.

A polynomial involving more than one variable can also be represented in the form of Eq. (11.1.11). For example, suppose that the values of four variables r, s, t, and y are observed for several different patients, and we wish to fit to these observed values a function having the following form:

$$y = \beta_0 + \beta_1 r + \beta_2 r^2 + \beta_3 rs + \beta_4 s^2 + \beta_5 t^3 + \beta_6 rst. \tag{11.1.16}$$

We can regard the function in Eq. (11.1.16) as a linear function having the form given in Eq. (11.1.11) with $k = 6$ if we define the six variables $x_1, \dots, x_6$ as follows: $x_1 = r$, $x_2 = r^2$, $x_3 = rs$, $x_4 = s^2$, $x_5 = t^3$, and $x_6 = rst$.

Summary

The method of least squares allows the calculation of a predictor for one variable (y) based on one or more other variables ($x_1, \ldots, x_k$) of the form $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. The coefficients $\beta_0, \ldots, \beta_k$ are chosen so that the sum of squared differences between observed values of y and observed values of $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ is as small as possible. Algebraic formulas for the coefficients are given for the case $k = 1$, but most statistical computer software will calculate the coefficients more easily.

Exercises

1. Prove that $\sum_{i=1}^n (c_1 x_i + c_2)^2 = c_1^2 \sum_{i=1}^n (x_i - \overline{x})^2 + n(c_1\overline{x} + c_2)^2$.

2. Show that the value of $\hat{\beta}_1$ in Eq. (11.1.1) can be rewritten in each of the following three forms:

a. $\hat{\beta}_1 = \dfrac{\sum_{i=1}^n x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}$

b. $\hat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}$

c. $\hat{\beta}_1 = \dfrac{\sum_{i=1}^n x_i (y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$

3. Show that the least-squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ passes through the point $(\overline{x}, \overline{y})$.

4. For $i = 1, \ldots, n$, let $\hat{y}_i = \beta_0 + \beta_1 x_i$. Show that $\hat{\beta}_0$ and $\hat{\beta}_1$, as given by Eq. (11.1.1), are the unique values of β_0 and β_1 such that

$$\sum_{i=1}^n (y_i - \hat{y}_i) = 0 \quad \text{and} \quad \sum_{i=1}^n x_i (y_i - \hat{y}_i) = 0.$$

5. Fit a straight line to the observed values given in Table 11.1 so that the sum of the squares of the *horizontal* deviations of the points from the line is a minimum. Sketch on the same graph both this line and the least-squares line given in Fig. 11.4.

6. Suppose that both the least-squares line and the least-squares parabola were fitted to the same set of points. Explain why the sum of the squares of the deviations of the points from the parabola cannot be larger than the sum of the squares of the deviations of the points from the straight line.

7. Suppose that eight specimens of a certain type of alloy were produced at different temperatures, and the durability of each specimen was then observed. The observed values are given in Table 11.3, where x_i denotes the temperature (in coded units) at which specimen i was produced and y_i denotes the durability (in coded units) of that specimen.

Table 11.3 Data for Exercise 7

i	x_i	y_i
1	0.5	40
2	1.0	41
3	1.5	43
4	2.0	42
5	2.5	44
6	3.0	42
7	3.5	43
8	4.0	42

a. Fit a straight line of the form $y = \beta_0 + \beta_1 x$ to these values by the method of least squares.

b. Fit a parabola of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2$ to these values by the method of least squares.

c. Sketch on the same graph the eight data points, the line found in part (a), and the parabola found in part (b).

8. Let (x_i, y_i) for $i = 1, \ldots, k + 1$, denote $k + 1$ given points in the xy-plane such that no two of these points have the same x-coordinate. Show that there is a unique polynomial having the form $y = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$ that passes through these $k + 1$ points.

9. The resilience y of a certain type of plastic is to be represented as a linear function of both the temperature x_1 at which the plastic is baked and the number of minutes x_2 for which it is baked. Suppose that 10 pieces of plastic are prepared by using different values of x_1 and x_2, and the observed values in appropriate units are as given in Table 11.4. Fit a function having the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ to these observed values by the method of least squares.

10. Consider again the observed values presented in Table 11.4. Fit a function having the form $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$ to these values by the method of least squares.

Table 11.4 Data for Exercise 9

i	x_{i1}	x_{i2}	y_i	i	x_{i1}	x_{i2}	y_i
1	100	1	113	6	120	2	144
2	100	2	118	7	120	3	138
3	110	1	127	8	130	1	146
4	110	2	132	9	130	2	156
5	120	1	136	10	130	3	149

11. Consider again the observed values presented in Table 11.4, and consider also the two functions that were fitted to these values in Exercises 9 and 10. Which of these two functions fits the observed values better?

11.2 Regression

In Sec. 11.1, we introduced the method of least squares. This method computes coefficients for a linear function to predict one variable y based on other variables $x_1, \dots, x_k$. In this section, we assume that the y values are observed values of a collection of random variables. In this case, there is a statistical model in which the method of least squares turns out to produce the maximum likelihood estimates of the parameters of the model.

Regression Functions

Example 11.2.1 **Pressure and the Boiling Point of Water.** Forbes (1857) reports the results from experiments that were trying to obtain a method for estimating altitude. A formula is available for altitude in terms of barometric pressure, but it was difficult to carry a barometer to high altitudes in Forbes' day. However, it might be easy for travelers to carry a thermometer and measure the boiling point of water. Table 11.5 contains the measured barometric pressures and boiling points of water from 17 experiments. We can use the method of least squares to fit a linear relationship between boiling point and pressure. Let y_i be the pressure for one of Forbes' observations, and let x_i be the corresponding boiling point for $i = 1, \dots, 17$. Using the data in Table 11.5, we can compute the least-squares line. The intercept and slope are, respectively, $\hat{\beta}_0 = -81.049$ and $\hat{\beta}_1 = 0.5228$. Of course, we do not expect that the line $y = -81.049 + 0.5228x$ precisely gives the relationship between boiling point x and pressure y. If we learn the boiling point x of water and want to compute the conditional distribution of the unknown pressure Y, is there a statistical model that allows us to say what the (conditional) distribution of pressure is given that the boiling point is x? ◄

In this section, we shall describe a statistical model for problems such as the one in Example 11.2.1. Fitting this statistical model will make use of the method of least squares. We shall study problems in which we are interested in learning about the conditional distribution of some random variable Y for given values of some other variables $X_1, \dots, X_k$. The variables $X_1, \dots, X_k$ may be random variables whose values are to be observed in an experiment along with the values of Y, or they may be *control variables* whose values are to be chosen by the experimenter. In general, some

Table 11.5 Boiling point of water in degrees Fahrenheit and atmospheric pressure in inches of mercury from Forbes' experiments. These data are taken from Weisberg (1985, p. 3).

Boiling Point	Pressure
194.5	20.79
194.3	20.79
197.9	22.40
198.4	22.67
199.4	23.15
199.9	23.35
200.9	23.89
201.1	23.99
201.4	24.02
201.3	24.01
203.6	25.14
204.6	26.57
209.5	28.49
208.6	27.76
210.7	29.04
211.9	29.88
212.2	30.06

of these variables might be random variables, and some might be control variables. In any case, we can study the conditional distribution of Y given $X_1, \dots, X_k$. We begin with some terminology.

Definition 11.2.1 Response/Predictor/Regression. The variables $X_1, \dots, X_k$ are called *predictors*, and the random variable Y is called the *response*. The conditional expectation of Y for given values $x_1, \dots, x_k$ of $X_1, \dots, X_k$ is called the *regression function of Y on $X_1, \dots, X_k$*, or simply the *regression of Y on $X_1, \dots, X_k$*.

The regression of Y on $X_1, \dots, X_k$ is a function of the values $x_1, \dots, x_k$ of $X_1, \dots, X_k$. In symbols, this function is $E(Y|x_1, \dots, x_k)$.

In this chapter, we shall assume that the regression function $E(Y|x_1, \dots, x_k)$ is a linear function having the following form:

$$E(Y|x_1, \dots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{11.2.1}$$

The coefficients $\beta_0, \dots, \beta_k$ in Eq. (11.2.1) are called *regression coefficients*. We shall suppose that these regression coefficients are unknown. Therefore, they are to be regarded as parameters whose values are to be estimated. We shall suppose also that n vectors of observations are obtained. For $i = 1, \dots, n$, we shall assume that the ith vector $(x_{i1}, \dots, x_{ik}, y_i)$ consists of a set of controlled or observed values of $X_1, \dots, X_k$ and the corresponding observed value of Y.

One set of estimators of the regression coefficients $\beta_0, \ldots, \beta_k$ that can be calculated from these observations is the set of values $\hat{\beta}_0, \ldots, \hat{\beta}_k$ that are obtained by the method of least squares, as described in Sec. 11.1. These estimators are called the *least-squares estimators* of $\beta_0, \ldots, \beta_k$. We shall now specify some further assumptions about the conditional distribution of Y given $X_1, \ldots, X_k$ in order to be able to determine in greater detail the properties of these least-squares estimators.

Simple Linear Regression

We shall consider first a problem in which we wish to study the regression of Y on just a single variable X. We shall assume that for each value $X = x$, the random variable Y can be represented in the form $Y = \beta_0 + \beta_1 x + \varepsilon$, where ε is a random variable that has the normal distribution with mean 0 and variance σ^2. It follows from this assumption that the conditional distribution of Y given $X = x$ is the normal distribution with mean $\beta_0 + \beta_1 x$ and variance σ^2.

A problem of this type is called a problem of *simple linear regression*. Here the term *simple* refers to the fact that we are considering the regression of Y on just a single variable X, rather than on more than one variable; the term *linear* refers to the fact that the regression function $E(Y|x) = \beta_0 + \beta_1 x$ is a linear function of the parameters β_0 and β_1. For example, a problem in which $E(Y|x)$ is a polynomial, like the right side of Eq. (11.1.6), would also be a linear regression problem, but not simple.

Throughout this section (and the next two sections), we shall consider the problem in which we shall observe n pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$. We shall make the following five assumptions. Each of these assumptions has a natural generalization to the case in which there is more than one predictor, but we shall postpone discussion of that case until Sec. 11.5.

Assumption 11.2.1 Predictor is known. Either the values $x_1, \ldots, x_n$ are known ahead of time or they are the observed values of random variables $X_1, \ldots, X_n$ on whose values we condition before computing the joint distribution of $(Y_1, \ldots, Y_n)$.

Assumption 11.2.2 Normality. For $i = 1, \ldots, n$, the conditional distribution of Y_i given the values $x_1, \ldots, x_n$ is a normal distribution.

Assumption 11.2.3 Linear Mean. There are parameters β_0 and β_1 such that the conditional mean of Y_i given the values $x_1, \ldots, x_n$ has the form $\beta_0 + \beta_1 x_i$ for $i = 1, \ldots, n$.

Assumption 11.2.4 Common Variance. There is a parameter σ^2 such that the conditional variance of Y_i given the values $x_1, \ldots, x_n$ is σ^2 for $i = 1, \ldots, n$. This assumption is often called *homoscedasticity*. Random variables with different variances are called *heteroscedastic*.

Assumption 11.2.5 Independence. The random variables $Y_1, \ldots, Y_n$ are independent given the observed $x_1, \ldots, x_n$.

A brief word is in order about Assumption 11.2.1. In Example 11.1.1, we saw that the reaction x_i of patient i to standard drug A is observed as part of the experiment along with the reaction y_i to drug B. Hence, the predictors are not known in advance. In this case, all probability statements that we make in this example are conditional on $(x_1, \ldots, x_n)$. In other examples, one might be trying to predict an economic variable using the year in which it was measured. In such cases, such as Example 11.5.1, which

we will see later, the values of at least some of the predictors are truely known in advance.

Assumptions 11.2.1–11.2.5 specify the conditional joint distribution of $Y_1, \ldots, Y_n$ given the vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ and the parameters β_0, β_1, and σ^2. In particular, the conditional joint p.d.f. of $Y_1, \ldots, Y_n$ is

$$f_n(\boldsymbol{y}|\boldsymbol{x}, \beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]. \qquad (11.2.2)$$

We can now find the M.L.E.'s of β_0, β_1, and σ^2.

Theorem 11.2.1 Simple Linear Regression M.L.E.'s. Assume Assumptions 11.2.1–11.2.5. The M.L.E.'s of β_0 and β_1 are the least-squares estimates, and the M.L.E. of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \qquad (11.2.3)$$

Proof For each observed vector $\boldsymbol{y} = (y_1, \ldots, y_n)$, the p.d.f. (11.2.2) will be the likelihood function of the parameters β_0, β_1, and σ^2. In Eq. (11.2.2), β_0 and β_1 appear only in the sum of squares

$$Q = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2,$$

which in turn appears in the exponent multiplied by $-1/[2\sigma^2]$. Regardless of the value of σ^2, the exponent is maximized over β_0 and β_1 by minimizing Q. It follows that the M.L.E.'s can be found in sequence by first minimizing Q over β_0 and β_1, then inserting the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that provide the minimum of Q, and finally minimizing the result over σ^2. The reader will note that Q is the same as the sum of squares in Eq. (11.1.2), which is minimized by the method of least squares. Thus, the M.L.E.'s of the regression coefficients β_0 and β_1 are precisely the same as the least-squares estimates. The exact form of these estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ was given in Eq. (11.1.1).

To find the M.L.E. of σ^2, perform the the second and third steps described in the preceding paragraph, namely, first replace β_0 and β_1 in Eq. (11.2.2) by their M.L.E.'s $\hat{\beta}_0$ and $\hat{\beta}_1$, and then maximize the resulting expression with respect to σ^2. The details are left to Exercise 1 at the end of this section, and the result is (11.2.3). ∎

The Distribution of the Least-Squares Estimators

We shall now present the joint distribution of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ when they are regarded as functions of the random variables $Y_1, \ldots, Y_n$ for given values of $x_1, \ldots, x_n$. Specifically, the estimators are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{x},$$

where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

It is convenient, both for this section and the next, to introduce the symbol

$$s_x = \left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{1/2}. \qquad (11.2.4)$$

Theorem 11.2.2 Distributions of Least-Squares Estimators. Under Assumptions 11.2.1–11.2.5, the distribution of $\hat{\beta}_1$ is the normal distribution with mean β_1 and variance σ^2/s_x^2. The distribution of $\hat{\beta}_0$ is the normal distribution with mean β_0 and variance

$$\sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{s_x^2}\right). \tag{11.2.5}$$

Finally, the covariance of $\hat{\beta}_1$ and $\hat{\beta}_0$ is

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\overline{x}\sigma^2}{s_x^2}. \tag{11.2.6}$$

(All of the distributional statements in this theorem are conditional on $X_i = x_i$ for $i = 1, \ldots, n$ if $X_1, \ldots, X_n$ are random variables.)

Proof To determine the distribution of $\hat{\beta}_1$, it is convenient to write $\hat{\beta}_1$ as follows (see Exercise 2 at the end of Sec. 11.1):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{s_x^2}. \tag{11.2.7}$$

It can be seen from Eq. (11.2.7) that $\hat{\beta}_1$ is a linear function of $Y_1, \ldots, Y_n$. Because the random variables $Y_1, \ldots, Y_n$ are independent and each has a normal distribution, it follows that $\hat{\beta}_1$ will also have a normal distribution. Furthermore, the mean of this distribution will be

$$E(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \overline{x})E(Y_i)}{s_x^2}.$$

Because $E(Y_i) = \beta_0 + \beta_1 x_i$ for $i = 1, \ldots, n$, it can now be found (see Exercise 2 at the end of this section) that

$$E(\hat{\beta}_1) = \beta_1. \tag{11.2.8}$$

Furthermore, because the random variables $Y_1, \ldots, Y_n$ are independent and each has variance σ^2, it follows from Eq. (11.2.7) that

$$\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \overline{x})^2 \,\text{Var}(Y_i)}{s_x^4} = \frac{\sigma^2}{s_x^2}. \tag{11.2.9}$$

Next, consider the distribution of $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{x}$. Because both $\overline{Y}$ and $\hat{\beta}_1$ are linear functions of $Y_1, \ldots, Y_n$, it follows that $\hat{\beta}_0$ is also a linear function of $Y_1, \ldots, Y_n$. Hence, $\hat{\beta}_0$ will have a normal distribution. The mean of $\hat{\beta}_0$ can be determined from the relation $E(\hat{\beta}_0) = E(\overline{Y}) - \overline{x}E(\hat{\beta}_1)$. It can be shown (see Exercise 3) that $E(\hat{\beta}_0) = \beta_0$. Furthermore, it can be shown (see Exercise 4) that $\text{Var}(\hat{\beta}_0)$ is given by (11.2.5). Finally, it can be shown (see Exercise 5) that the value of the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ is given by (11.2.6). ■

A simple corollary to Theorem 11.2.2 is that $\hat{\beta}_0$ and $\hat{\beta}_1$ are, respectively, unbiased estimators of the corresponding parameters β_0 and β_1.

To complete the description of the joint distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$, it will be shown in Sec. 11.3 that this joint distribution is the bivariate normal distribution for which the means, variances, and covariance are as stated in Theorem 11.2.2.

Example 11.2.2 Pressure and the Boiling Point of Water. In Example 11.2.1, we found the least-squares line for predicting pressure from boiling point of water. Suppose that we use the linear regression model just described as a model for the data in this experiment. That is, let Y_i be the pressure for one of Forbes' observations, and let x_i be the corresponding boiling point for $i = 1, \dots, 17$. We model the Y_i as being independent with means $\beta_0 + \beta_1 x_i$ and variance σ^2. The average temperature is $\overline{x} = 202.95$ and $s_x^2 = 530.78$ with $n = 17$. From these values, we can now compute the variances and covariances of the least-squares estimators using the formulas derived in this section. For example,

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{530.78} = 0.00188\sigma^2,$$

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{17} + \frac{202.95^2}{530.78}\right) = 77.66\sigma^2,$$

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{202.95\sigma^2}{530.78} = 0.382\sigma^2.$$

It is easy to see that we expect to get a much more precise estimate of β_1 than of β_0. ◀

The statement at the end of Example 11.2.2 about getting more precise estimates of β_1 than of β_0 is a bit deceptive. We must multiply β_1 by a number on the order of 200 before it is on the same scale as β_0. Hence, it might make more sense to compare the variance of $200\hat{\beta}_1$ to the variance of $\hat{\beta}_0$. In general, we can find the variance of any linear combination of the least-squares estimators.

Example 11.2.3 The Variance of a Linear Combination. Very often, we need to compute the variance of a linear combination of the least-squares estimators. One example is prediction, as discussed later in this section. Suppose that we wish to compute the variance of $T = c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + c_*$. The variance of T can be found by substituting the values of $\mathrm{Var}(\hat{\beta}_0)$, $\mathrm{Var}(\hat{\beta}_1)$, and $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ given in Eqs. (11.2.5), (11.2.9), and (11.2.6) in the following relation:

$$\mathrm{Var}(T) = c_0^2\,\mathrm{Var}(\hat{\beta}_0) + c_1^2\,\mathrm{Var}(\hat{\beta}_1) + 2c_0c_1\,\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1).$$

When these substitutions have been made, the result can be written in the following form:

$$\mathrm{Var}(T) = \sigma^2\left(\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2}\right). \tag{11.2.10}$$

For the specific case of Example 11.2.2, we have $c_0 = 0$ and $c_1 = 200$, so the variance of $200\hat{\beta}_1$ is $200^2\sigma^2/s_x^2 = 75.36\sigma^2$. This is pretty close to the variance of $\hat{\beta}_0$, namely, $77.66\sigma^2$. ◀

Prediction

Example 11.2.4 Predicting Pressure from the Boiling Point of Water. In Example 11.2.1, Forbes was trying to find a way to use the boiling point of water to estimate the barometric pressure. Suppose that a traveler measures the boiling point of water to be 201.5 degrees. What estimate of barometric pressure should they give and how much uncertainty is there about this estimate? ◀

Suppose that n pairs of observations $(x_1, Y_1), \dots, (x_n, Y_n)$ are to be obtained in a problem of simple linear regression, and on the basis of these n pairs, it is necessary to predict the value of an independent observation Y that will be obtained when a certain specified value x is assigned to the control variable. Since the observation Y will have the normal distribution with mean $\beta_0 + \beta_1 x$ and variance σ^2, it is natural to use the value $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ as the predicted value of Y. We shall now determine the M.S.E. $E[(\hat{Y} - Y)^2]$ of this prediction, where both $\hat{Y}$ and Y are random variables.

Theorem 11.2.3 M.S.E. of Prediction. In the prediction problem just described,

$$E[(\hat{Y} - Y)^2] = \sigma^2\left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{s_x^2}\right]. \tag{11.2.11}$$

Proof In this problem, $E(\hat{Y}) = E(Y) = \beta_0 + \beta_1 x$. Thus, if we let $\mu = \beta_0 + \beta_1 x$, then

$$\begin{aligned} E[(\hat{Y} - Y)^2] &= E\{[(\hat{Y} - \mu) - (Y - \mu)]^2\} \\ &= \mathrm{Var}(\hat{Y}) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(\hat{Y}, Y). \end{aligned} \tag{11.2.12}$$

However, the random variables $\hat{Y}$ and Y are independent, because $\hat{Y}$ is a function of the first n pairs of observations and Y is an independent observation. Therefore, $\mathrm{Cov}(\hat{Y}, Y) = 0$, and it follows that

$$E[(\hat{Y} - Y)^2] = \mathrm{Var}(\hat{Y}) + \mathrm{Var}(Y). \tag{11.2.13}$$

Finally, because $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$, the value of $\mathrm{Var}(\hat{Y})$ is given by Eq. (11.2.10) with $c_0 = 1$ and $c_1 = x$. Also $\mathrm{Var}(Y) = \sigma^2$. Substituting these into Eq. (11.2.13) gives (11.2.11). ■

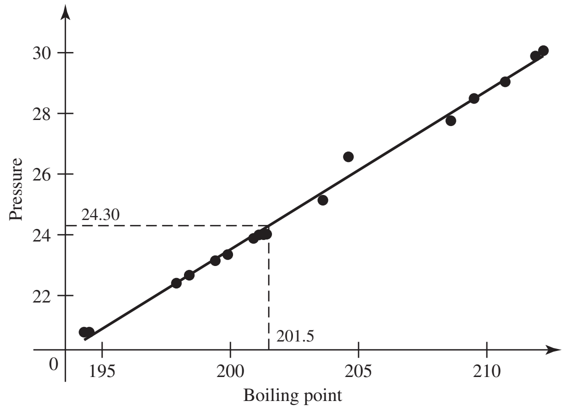
Example 11.2.5 Predicting Pressure from the Boiling Point of Water. In Example 11.2.4, we wanted to predict barometric pressure when the boiling point of water is 201.5 degrees. The least-squares line is $y = -81.049 + 0.5228x$, and $\hat{\sigma}^2 = 0.0478$. Fig. 11.7 shows the data plotted together with the least-squares regression line and the location of the point on the line that has $x = 201.5$. The M.S.E. of the prediction of pressure Y is obtained from Eq. (11.2.11):

$$E[(\hat{Y} - Y)^2] = \sigma^2\left[1 + \frac{1}{17} + \frac{(201.5 - 202.95)^2}{530.78}\right] = 1.0628\sigma^2,$$

and the observed value of the prediction is $\hat{Y} = -81.06 + 0.5229 \times 201.5 = 24.30$. The calculation of $\hat{Y}$ is illustrated in Fig. 11.7. The M.S.E. $1.0628\sigma^2$ can be interpreted as follows: If we knew the values of β_0 and β_1 and tried to predict Y, the M.S.E. would be $\mathrm{Var}(Y) = \sigma^2$. Having to estimate β_0 and β_1 only costs us an additional $0.0628\sigma^2$ in M.S.E. ◄

Note: M.S.E. of Prediction Increases as x Moves Away from Observed Data. The M.S.E. in Eq. (11.2.11) increases as x moves away from $\overline{x}$, and it is smallest when $x = \overline{x}$. This indicates that it is harder to predict Y when x is not near the center of the observed values $x_1, \dots, x_n$. Indeed, if x is larger than the largest observed x_i or smaller than the smallest one, it is quite difficult to predict Y with much precision. Such predictions outside the range of the observed data are called *extrapolations*.

Figure 11.7 Plot of pressure versus boiling point with regression line for Example 11.2.5. Dotted line illustrates prediction of pressure when boiling point is 201.5.



Design of the Experiment

Consider a problem of simple linear regression in which the variable X is a control variable whose values $x_1, \ldots, x_n$ can be chosen by the experimenter. We shall discuss methods for choosing these values so as to obtain good estimators of the regression coefficients β_0 and β_1.

Suppose first that the values $x_1, \ldots, x_n$ are to be chosen so as to minimize the M.S.E. of the least-squares estimator $\hat{\beta}_0$. Since $\hat{\beta}_0$ is an unbiased estimator of β_0, the M.S.E. of $\hat{\beta}_0$ is equal to $\text{Var}(\hat{\beta}_0)$, as given in Eq. (11.2.5). It follows from Eq. (11.2.5) that $\text{Var}(\hat{\beta}_0) \geq \sigma^2/n$ for all values $x_1, \ldots, x_n$, and there will be equality in this relation if and only if $\overline{x} = 0$. Hence, $\text{Var}(\hat{\beta}_0)$ will attain its minimum value σ^2/n whenever $\overline{x} = 0$. Of course, this will be impossible in any application in which X is constrained to be positive.

Suppose next that the values $x_1, \ldots, x_n$ are to be chosen so as to minimize the M.S.E. of the estimator $\hat{\beta}_1$. Again, the M.S.E. of $\hat{\beta}_1$ will be equal to $\text{Var}(\hat{\beta}_1)$, as given in Eq. (11.2.9). It can be seen from Eq. (11.2.9) that $\text{Var}(\hat{\beta}_1)$ will be minimized by choosing the values $x_1, \ldots, x_n$ so that the value of s_x^2 is maximized. If the values $x_1, \ldots, x_n$ must be chosen from some bounded interval (a, b) of the real line, and if n is an even integer, then the value of s_x^2 will be maximized by choosing $x_i = a$ for exactly $n/2$ values and choosing $x_i = b$ for the other $n/2$ values. If n is an odd integer, all the values should again be chosen at the endpoints a and b, but one endpoint must now receive one more observation than the other endpoint.

It follows from this discussion that if the experiment is to be designed so as to minimize both the M.S.E. of $\hat{\beta}_0$ and the M.S.E. of $\hat{\beta}_1$, then the values $x_1, \ldots, x_n$ should be chosen so that exactly, or approximately, $n/2$ values are equal to some number c that is as large as is feasible in the given experiment, and the remaining values are equal to $-c$. In this way, the value of $\overline{x}$ will be exactly, or approximately, equal to 0, and the value of s_x^2 will be as large as possible.

Finally, suppose that the linear combination $\theta = c_0\beta_0 + c_1\beta_1 + c_*$ is to be estimated, where $c_0 \neq 0$, and that the experiment is to be designed so as to minimize the M.S.E. of $\hat{\theta}$, that is, to minimize $\text{Var}(\hat{\theta})$. For example, if Y is a future observation with corresponding predictor x, then we could set $c_0 = 1$, $c_2 = x$, and $c_* = 0$ in order to make $\theta = E(Y|x)$. In Example 11.2.3, we computed $\text{Var}(T)$, where $T = \hat{\theta}$, as the sum of two nonnegative terms in Eq. (11.2.10). The second term is the only one that

depends on the values of $x_1, \ldots, x_n$, and it equals 0 (its smallest possible value) if and only if $\overline{x} = c_1/c_0$. In this case, $\text{Var}(\hat{\theta})$ will attain its minimum value $c_0^2\sigma^2/n$.

In practice, an experienced statistician would not usually choose all the values $x_1, \ldots, x_n$ at a single point or at just the two endpoints of the interval (a, b), as the optimal designs that we have just derived would dictate. The reason is that when all n observations are taken at just one or two values of X, the experiment provides no possibility of checking the assumption that the regression of Y on X is a linear function. In order to check this assumption without unduly increasing the M.S.E. of the least-squares estimators, many of the values $x_1, \ldots, x_n$ should be chosen at the endpoints a and b, but at least some of the values should be chosen at a few interior points of the interval. Linearity can then be checked by visual inspection of the plotted points and the fitting of a polynomial of degree two or higher.

Summary

We considered the following statistical model. The values $x_1, \ldots, x_n$ are assumed known. The random variables $Y_1, \ldots, Y_n$ are independent with Y_i having the normal distribution with mean $\beta_0 + \beta_1 x_i$ and variance σ^2. Here, β_0, β_1, and σ^2 are unknown parameters. These are the assumptions of the simple linear regression model. Under this model, the joint distribution of the least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ is a bivariate normal distribution with $\hat{\beta}_i$ having mean β_i for $i = 1, 2$. The variances are given in Eqs. (11.2.5) and (11.2.9). The covariance is given in Eq. (11.2.6). If we consider predicting a future Y value with corresponding predictor x, we might use the prediction $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. In this case, $Y - \hat{Y}$ has the normal distribution with mean 0 and variance given by Eq. (11.2.11).

Exercises

1. Show that the M.L.E. of σ^2 is given by Eq. (11.2.3).

2. Show that $E(\hat{\beta}_1) = \beta_1$.

3. Show that $E(\hat{\beta}_0) = \beta_0$.

4. Show that $\text{Var}(\hat{\beta}_0)$ is as given in Eq. (11.2.5).

5. Show that $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ is as given in Eq. (11.2.6). *Hint:* Use the result in Exercise 8 in Sec. 4.6.

6. Show that in a problem of simple linear regression, the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ will be independent if $\overline{x} = 0$.

7. Consider a problem of simple linear regression in which a patient's reaction Y to a new drug B is to be related to his reaction X to a standard drug A. Suppose that the 10 pairs of observed values given in Table 11.1 are obtained.

a. Determine the values of the M.L.E.'s $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$.

b. Determine the values of $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1)$.

c. Determine the value of the correlation of $\hat{\beta}_0$ and $\hat{\beta}_1$.

8. Consider again the conditions of Exercise 7, and suppose that it is desired to estimate the value of $\theta = 3\beta_0 - 2\beta_1 + 5$. Determine an unbiased estimator of θ and find its M.S.E.

9. Consider again the conditions of Exercise 7, and let $\theta = 3\beta_0 + c_1\beta_1$, where c_1 is a constant. Determine an unbiased estimator $\hat{\theta}$ of θ. For what value of c_1 will the M.S.E. of $\hat{\theta}$ be smallest?

10. Consider again the conditions of Exercise 7. If a particular patient's reaction to drug A has the value $x = 2$, what is the predicted value of his reaction to drug B, and what is the M.S.E. of this prediction?

11. Consider again the conditions of Exercise 7. For what value x of a patient's reaction to drug A can his reaction to drug B be predicted with the smallest M.S.E.?

12. Consider a problem of simple linear regression in which the durability Y of a certain type of alloy is to be related to the temperature X at which it was produced. Suppose that the eight pairs of observed values given in Table 11.3 are obtained. Determine the values of the M.L.E.'s $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$, and also the values of $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1)$.

13. For the conditions of Exercise 12, determine the value of the correlation of $\hat{\beta}_0$ and $\hat{\beta}_1$.

14. Consider again the conditions of Exercise 12, and suppose that it is desired to estimate the value of $\theta = 5 - 4\beta_0 + \beta_1$. Find an unbiased estimator $\hat{\theta}$ of θ. Determine the value of $\hat{\theta}$ and the M.S.E. of $\hat{\theta}$.

15. Consider again the conditions of Exercise 12, and let $\theta = c_1\beta_1 - \beta_0$, where c_1 is a constant. Determine an unbiased estimator $\hat{\theta}$ of θ. For what value of c_1 will the M.S.E. of $\hat{\theta}$ be smallest?

16. Consider again the conditions of Exercise 12. If a specimen of the alloy is to be produced at the temperature $x = 3.25$, what is the predicted value of the durability of the specimen, and what is the M.S.E. of this prediction?

17. Consider again the conditions of Exercise 12. For what value of the temperature x can the durability of a specimen of the alloy be predicted with the smallest M.S.E.?

18. Moore and McCabe (1999, p. 174) report prices paid for several species of seafood in 1970 and 1980. These values are in Table 11.6. If we were interested in trying to predict 1980 seafood prices from 1970 prices, a linear regression model might be used.

a. Find the least-squares regression coefficients for predicting 1980 prices from 1970 prices.

b. If an additional species sold for 21.4 in 1970, what would you predict for the 1980 selling price?

c. What is the M.S.E. for predicting the 1980 price of a species that sold for 21.4 in 1970?

Table 11.6 Fish prices in 1970 and 1980 for Exercise 18

1970	1980	1970	1980
13.1	27.3	26.7	80.1
15.3	42.4	47.5	150.7
25.8	38.7	6.6	20.3
1.8	4.5	94.7	189.7
4.9	23	61.1	131.3
55.4	166.3	135.6	404.2
39.3	109.7	47.6	149

19. In the 1880s, Francis Galton studied the inheritance of physical characteristics. Galton found that the sons of tall men tended to be taller than average, but shorter than their fathers. Similarly, sons of short men tended to be shorter than average, but taller than their fathers. Thus, the average heights of the sons were closer to the mean height of the population, regardless of whether the fathers were taller or shorter than average. From these observations, one might conclude that the variability of height decreases over successive generations, both tall persons and short persons tend to be eliminated, and the population "regresses" toward some average height. This conclusion is an example of the *regression fallacy*. In this problem you will prove that the regression fallacy arises in the bivariate normal distribution even when both coordinates have the same variance. In particular, assume that the vector (X_1, X_2) has the bivariate normal distribution with common mean μ, common variance σ^2, and positive correlation $\rho < 1$. Prove that $E(X_2|x_1)$ is closer to μ than x_1 is to μ for every value x_1. (This occurs despite the fact that X_1 and X_2 have the same mean and the same variance.)

11.3 Statistical Inference in Simple Linear Regression

Many of the inference procedures introduced in Chapters 8 and 9 that were used for samples from a normal distribution can be extended to the simple linear regression model. The theorems that allowed us to conclude that various statistics had t distributions will continue to apply in the regression case.

Joint Distribution of the Estimators

Example 11.3.1 **Pressure and the Boiling Point of Water.** Consider the traveler in Example 11.2.4, who is interested in the barometric pressure when the boiling point of water is 201.5 degrees. Suppose that this traveler would like to know whether the pressure is 24.5. For example, the traveler might wish to test the null hypothesis $H_0: \beta_0 + 201.5\beta_1 = 24.5$.

Alternatively, the traveler might desire an interval estimate of $\beta_0 + 201.5\beta_1$. Such inferences are possible once we find the joint distribution of the estimators of all of the parameters (β_0, β_1, and σ^2) of the regression model. ◄

It was stated after the proof of Theorem 11.2.2 that, in a problem of simple linear regression, the joint distribution of the M.L.E.'s $\hat{\beta}_0$ and $\hat{\beta}_1$ is the bivariate normal distribution for which the means, the variances, and the covariance are specified in Theorem 11.2.2. In this section, we shall prove this fact. We shall also consider the M.L.E. $\hat{\sigma}^2$, which was presented in Eq. (11.2.3), and we shall derive the joint distribution of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$. In particular, we shall show that the estimator $\hat{\sigma}^2$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$.

We continue to make Assumptions 11.2.1–11.2.5. The derivation of the joint distribution of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$, which we shall present, is based on the properties of orthogonal matrices, as described in Sec. 8.3.

We shall continue to use the definition of s_x in Eq. (11.2.4). Also, let $\boldsymbol{a}_1 = (a_{11}, \ldots, a_{1n})$ and $\boldsymbol{a}_2 = (a_{21}, \ldots, a_{2n})$ be n-dimensional vectors, which are defined as follows:

$$a_{1j} = \frac{1}{n^{1/2}} \quad \text{for } j = 1, \ldots, n, \tag{11.3.1}$$

and

$$a_{2j} = \frac{1}{s_x}(x_j - \overline{x}) \quad \text{for } j = 1, \ldots, n. \tag{11.3.2}$$

It is easily verified that $\sum_{j=1}^{n} a_{1j}^2 = 1$, $\sum_{j=1}^{n} a_{2j}^2 = 1$, and $\sum_{j=1} a_{1j}a_{2j} = 0$.

Because the vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ have these properties, it is possible to construct an $n \times n$ orthogonal matrix $\boldsymbol{A}$ such that the coordinates of $\boldsymbol{a}_1$ form the first row of $\boldsymbol{A}$, and coordinates of $\boldsymbol{a}_2$ form the second row of $\boldsymbol{A}$. (To see how this is done, consult a linear algebra text, such as Cullen, 1972, p. 162, for the *Gram-Schmidt* method.) We shall assume that such a matrix $\boldsymbol{A}$ has been constructed:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}.$$

We shall now define a new random vector $\boldsymbol{Z}$ by the relation $\boldsymbol{Z} = \boldsymbol{AY}$, where

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{Z} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix}.$$

The joint distribution of $Z_1, \ldots, Z_n$ can be found from the following theorem, which is an extension of Theorem 8.3.4.

Theorem 11.3.1 Suppose that the random variables $Y_1, \ldots, Y_n$ are independent, and each has a normal distribution with the same variance σ^2. If $\boldsymbol{A}$ is an orthogonal $n \times n$ matrix and $\boldsymbol{Z} = \boldsymbol{AY}$, then the random variables $Z_1, \ldots, Z_n$ also are independent, and each has a normal distribution with variance σ^2.

Proof Let $E(Y_i) = \mu_i$ for $i = 1, \ldots, n$ (it is not assumed in the theorem that $Y_1, \ldots, Y_n$ have the same mean), and let

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}.$$

Also, let $\boldsymbol{X} = (1/\sigma)(\boldsymbol{Y} - \boldsymbol{\mu})$. Since it is assumed that the coordinates of the random vector $\boldsymbol{Y}$ are independent, then the coordinates of the random vector $\boldsymbol{X}$ will also be independent. Furthermore, each coordinate of $\boldsymbol{X}$ will have the standard normal distribution. Therefore, it follows from Theorem 8.3.4 that the coordinates of the n-dimensional random vector $\boldsymbol{AX}$ will also be independent, and each will have the standard normal distribution.

But

$$\boldsymbol{AX} = \frac{1}{\sigma}\boldsymbol{A}(\boldsymbol{Y} - \boldsymbol{\mu}) = \frac{1}{\sigma}\boldsymbol{Z} - \frac{1}{\sigma}\boldsymbol{A}\mu.$$

Hence,

$$\boldsymbol{Z} = \sigma\boldsymbol{AX} + \boldsymbol{A}\mu. \tag{11.3.3}$$

Since the coordinates of the random vector $\boldsymbol{AX}$ are independent, and each has the standard normal distribution, then the coordinates of the random vector $\sigma\boldsymbol{AX}$ will also be independent, and each will have the normal distribution with mean 0 and variance σ^2. When the vector $\boldsymbol{A}\mu$ is added to the random vector $\sigma\boldsymbol{AX}$, the mean of each coordinate will be shifted, but the coordinates will remain independent, and the variance of each coordinate will be unchanged. It now follows from Eq. (11.3.3) that the coordinates of the random vector $\boldsymbol{Z}$ will be independent, and each will have a normal distribution with variance σ^2. ■

In a problem of simple linear regression, the observations $Y_1, \ldots, Y_n$ satisfy the conditions of Theorem 11.3.1. Therefore, the coordinates of the random vector $\boldsymbol{Z} = \boldsymbol{AY}$ will be independent, and each will have a normal distribution with variance σ^2. We can use these facts to find the joint distribution of $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$.

Theorem 11.3.2 In the simple linear regression problem described above, the joint distribution of $(\hat{\beta}_0, \hat{\beta}_1)$ is the bivariate normal distribution for which the means, variances, and covariance are as stated in Theorem 11.2.2. Also, if $n \geq 3$, $\hat{\sigma}^2$ is independent of $(\hat{\beta}_0, \hat{\beta}_1)$ and $n\hat{\sigma}^2/\sigma^2$ has the χ^2 distribution with $n - 2$ degrees of freedom.

Proof The first two coordinates Z_1 and Z_2 of the random vector $\boldsymbol{Z}$ can easily be derived. The first coordinate is

$$Z_1 = \sum_{j=1}^{n} a_{1j}Y_j = \frac{1}{n^{1/2}}\sum_{j=1}^{n} Y_j = n^{1/2}\overline{Y}. \tag{11.3.4}$$

Since $\hat{\beta}_0 = \overline{Y} - \overline{x}\hat{\beta}_1$, we may also write

$$Z_1 = n^{1/2}(\hat{\beta}_0 + \overline{x}\hat{\beta}_1). \tag{11.3.5}$$

The second coordinate is

$$Z_2 = \sum_{j=1}^{n} a_{2j}Y_j = \frac{1}{s_x}\sum_{j=1}^{n}(x_j - \overline{x})Y_j. \tag{11.3.6}$$

By Eq. (11.2.7), we may also write

$$Z_2 = s_x\hat{\beta}_1. \tag{11.3.7}$$

Together, Eqs. (11.3.5) and (11.3.7) imply that

$$\begin{aligned}\hat{\beta}_0 &= n^{-1/2}Z_1 - \frac{\overline{x}}{s_x}Z_2,\\ \hat{\beta}_1 &= \frac{1}{s_x}Z_2.\end{aligned} \tag{11.3.8}$$

Since Z_1 and Z_2 are independent normal random variables, they have a bivariate normal joint distribution. Eqs. (11.3.8) express $\hat{\beta}_0$ and $\hat{\beta}_1$ as linear combinations of Z_1 and Z_2. These linear combinations satisfy the conditions of Exercise 10 of Sec. 5.10, which says in turn that $\hat{\beta}_0$ and $\hat{\beta}_1$ have a bivariate normal distribution. We already calculated the means, variances, and covariance in Theorem 11.2.2.

Now let the random variable S^2 be defined as follows:

$$S^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \tag{11.3.9}$$

(It is easy to see that the M.L.E. of σ^2, as given in Eq. (11.2.3), is $\hat{\sigma}^2 = S^2/n$.) We shall show that S^2 and the random vector $(\hat{\beta}_0, \hat{\beta}_1)$ are independent. Since $\hat{\beta}_0 = \overline{Y} - \overline{x}\hat{\beta}_1$, we may rewrite S^2 as follows:

$$\begin{aligned}S^2 &= \sum_{i=1}^{n}[Y_i - \overline{Y} - \hat{\beta}_1(x_i - \overline{x})]^2\\ &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - 2\hat{\beta}_1\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) + \hat{\beta}_1^2 s_x^2.\end{aligned}$$

It now follows from Eq. (11.1.1) that

$$S^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 - s_x^2\hat{\beta}_1^2. \tag{11.3.10}$$

Since $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{Y}$, where $\boldsymbol{A}$ is an orthogonal matrix, we know from Theorem 8.3.4 that $\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} Z_i^2$. By using this fact, we can now obtain the following relation from Eq. (11.3.4), (11.3.7), and (11.3.10):

$$S^2 = \sum_{i=1}^{n} Z_i^2 - Z_1^2 - Z_2^2 = \sum_{i=3}^{n} Z_i^2.$$

The random variables $Z_1, \ldots, Z_n$ are independent, and we have now shown that S^2 is equal to the sum of the squares of only $Z_3, \ldots, Z_n$. It follows, therefore, that S^2 and the random vector (Z_1, Z_2) are independent. But $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of Z_1 and Z_2 only, as seen in Eq. (11.3.8). Hence, S^2 and the random vector $(\hat{\beta}_0, \hat{\beta}_1)$ are independent.

We shall now derive the distribution of S^2. For $i = 3, \ldots, n$, we have $Z_i = \sum_{j=1}^{n} a_{ij}Y_j$. Hence,

$$
\begin{aligned}
E(Z_i) &= \sum_{j=1}^{n} a_{ij} E(Y_j) = \sum_{j=1}^{n} a_{ij}(\beta_0 + \beta_1 x_j) \\
&= \sum_{j=1}^{n} a_{ij}[\beta_0 + \beta_1 \overline{x} + \beta_1 (x_j - \overline{x})] \qquad (11.3.11) \\
&= (\beta_0 + \beta_1 \overline{x}) \sum_{j=1}^{n} a_{ij} + \beta_1 \sum_{j=1}^{n} a_{ij}(x_j - \overline{x}).
\end{aligned}
$$

Since the matrix $\boldsymbol{A}$ is orthogonal, the sum of the products of the corresponding terms in any two different rows must be 0. In particular, for $i = 3, \ldots, n$,

$$
\sum_{j=1}^{n} a_{ij} a_{1j} = 0 \quad \text{and} \quad \sum_{j=1}^{n} a_{ij} a_{2j} = 0.
$$

It now follows from the expressions for a_{1j} and a_{2j} given in Eqs. (11.3.1) and (11.3.2) that for $i = 3, \ldots, n$,

$$
\sum_{j=1}^{n} a_{ij} = 0 \quad \text{and} \quad \sum_{j=1}^{n} a_{ij}(x_j - \overline{x}) = 0.
$$

When these values are substituted into Eq. (11.3.11), it is found that $E(Z_i) = 0$ for $i = 3, \ldots, n$.

We now know that the $n - 2$ random variables $Z_3, \ldots, Z_n$ are independent, and that each has the normal distribution with mean 0 and variance σ^2. Since $S^2 = \sum_{i=3}^{n} Z_i^2$, it follows that the random variable S^2/σ^2 has the χ^2 distribution with $n - 2$ degrees of freedom.

Finally, we know that $\hat{\sigma}^2 = S^2/n$, and hence $\hat{\sigma}^2$ is independent of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, and the distribution of $n\hat{\sigma}^2/\sigma^2$ is the χ^2 distribution with $n - 2$ degrees of freedom. ■

Tests of Hypotheses about the Regression Coefficients

It will be convenient, for the remainder of the discussion of simple linear regression, to let

$$
\sigma' = \left(\frac{S^2}{n-2} \right)^{1/2}. \qquad (11.3.12)
$$

This random variable will appear in all of the test statistics and confidence intervals that we derive. It is analogous to the random variable with the same name that appears in Eqs. (8.4.3) and (8.4.5) and played a similar role in tests and confidence intervals for the mean of a single normal distribution.

We proved earlier that the joint distribution of $(\hat{\beta}_0, \hat{\beta}_1)$ is bivariate normal. This implies that every linear combination $c_0\hat{\beta}_0 + c_1\hat{\beta}_1$ has a normal distribution. We shall use this fact to simplify the discussion of inference about regression coefficients. We shall begin by deriving tests of hypotheses concerning a general linear combination $c_0\beta_0 + c_1\beta_1$ of the regression parameters. Then, specific cases will be introduced by choosing special values for c_0 and c_1. For example, $c_0 = 1$ and $c_1 = 0$ makes the linear combination β_0, while $c_0 = 0$ and $c_1 = 1$ leads to β_1.

Tests of Hypotheses about a Linear Combination of β_0 and β_1 Let c_0, c_1, and c_* be specified numbers, where at least one of c_0 and c_1 is nonzero, and suppose that we are interested in testing the following hypotheses:

$$\begin{aligned} H_0: &\quad c_0\beta_0 + c_1\beta_1 = c_*, \\ H_1: &\quad c_0\beta_0 + c_1\beta_1 \neq c_*. \end{aligned} \tag{11.3.13}$$

We shall derive a test of these hypotheses based on the random variables $c_0\hat{\beta}_0 + c_1\hat{\beta}_1$ and σ'.

Theorem 11.3.3 For each $0 < \alpha_0 < 1$, a level α_0 test of the hypotheses (11.3.13) is to reject H_0 if $|U_{01}| \geq T_{n-2}^{-1}(1 - \alpha_0/2)$, where

$$U_{01} = \left[\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2}\right]^{-1/2} \left(\frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*}{\sigma'}\right), \tag{11.3.14}$$

and T_{n-2}^{-1} is the quantile function of the t distribution with $n - 2$ degrees of freedom.

Proof In general, the mean of $c_0\hat{\beta}_0 + c_1\hat{\beta}_1$ is $c_0\beta_0 + c_1\beta_1$, and its variance was found in Eq. (11.2.10). Therefore, when H_0 is true, the following random variable W_{01} has the standard normal distribution:

$$W_{01} = \left[\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2}\right]^{-1/2} \left(\frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_*}{\sigma}\right).$$

Because the value of σ is unknown, a test of the hypotheses (11.3.13) cannot be based simply on the random variable W_{01}. However, the random variable S^2/σ^2 has the χ^2 distribution with $n - 2$ degrees of freedom for all possible values of the parameters β_0, β_1, and σ^2. Moreover, because $(\hat{\beta}_0, \hat{\beta}_1)$ is independent of S^2, it follows that W_{01} and S^2 are also independent. Hence, when the hypothesis H_0 is true, the random variable

$$\frac{W_{01}}{\left[\left(\frac{1}{n-2}\right)\left(\frac{S^2}{\sigma^2}\right)\right]^{1/2}} \tag{11.3.15}$$

has the t distribution with $n - 2$ degrees of freedom. It is straightforward to show that the expression in (11.3.15) also equals U_{01} in Eq. (11.3.14), which is a function of the observable data alone. It follows that the test specified in the theorem is a level α_0 test of the hypotheses (11.3.13). ■

The test procedure in Theorem 11.3.3 is also the likelihood ratio test procedure for the hypotheses (11.3.13), but the proof will not be given here.

Tests of One-Sided Hypotheses The same derivation just finished can also be used to form tests of hypotheses such as

$$\begin{aligned} H_0: &\quad c_0\beta_0 + c_1\beta_1 \leq c_*, \\ H_1: &\quad c_0\beta_0 + c_1\beta_1 > c_*, \end{aligned} \tag{11.3.16}$$

or

$$\begin{aligned} H_0: &\quad c_0\beta_0 + c_1\beta_1 \geq c_*, \\ H_1: &\quad c_0\beta_0 + c_1\beta_1 < c_*. \end{aligned} \tag{11.3.17}$$

The proof of the following result is similar to the proof of Theorem 11.3.3 and will not be given here.

Theorem 11.3.4 A level α_0 test of (11.3.16) is to reject H_0 if $U_{01} \geq T_{n-2}^{-1}(1-\alpha_0)$. A level α_0 test of (11.3.17) is to reject H_0 if $U_{01} \leq -T_{n-2}^{-1}(1-\alpha_0)$. ■

The only part of the proof of Theorem 11.3.4 that differs significantly from the corresponding part of Theorem 11.3.3 is the proof that the tests actually have level of significance α_0. The proof of this is similar to the proof of Theorem 9.5.1 and is left to the reader in Exercise 23.

We shall next present examples of how to test several common hypotheses concerning β_0 and β_1 by making use of the fact that U_{01} in Eq. (11.3.14) has the t distribution with $n-2$ degrees of freedom. These examples will correspond to setting c_0, c_1, and c_* equal to specific values.

Tests of Hypotheses about β_0 Let β_0^* be a specified number ($-\infty < \beta_0^* < \infty$), and suppose that it is desired to test the following hypotheses about the regression coefficient β_0:

$$\begin{aligned} H_0&: \quad \beta_0 = \beta_0^*, \\ H_1&: \quad \beta_0 \neq \beta_0^*. \end{aligned} \tag{11.3.18}$$

These hypotheses are the same as those in Eq. (11.3.13) if we make the substitutions $c_0 = 1$, $c_1 = 0$, and $c_* = \beta_0^*$. If we substitute these values into the formula for U_{01} in Eq. (11.3.14), we obtain the following random variable, U_0,

$$U_0 = \frac{\hat{\beta}_0 - \beta_0^*}{\sigma' \left[\dfrac{1}{n} + \dfrac{\overline{x}^2}{s_x^2} \right]^{1/2}}, \tag{11.3.19}$$

which then has the t distribution with $n-2$ degrees of freedom if H_0 is true.

Suppose that in a problem of simple linear regression, we are interested in testing the null hypothesis that the regression line $y = \beta_0 + \beta_1 x$ passes through the origin against the alternative hypothesis that the line does not pass through the origin. These hypotheses can be stated in the following form:

$$\begin{aligned} H_0&: \quad \beta_0 = 0, \\ H_1&: \quad \beta_0 \neq 0. \end{aligned} \tag{11.3.20}$$

Here the hypothesized value β_0^* is 0.

Let u_0 denote the value of U_0 calculated from a given set of observed values (x_i, y_i) for $i = 1, \ldots, n$. Then the tail area (p-value) corresponding to this value is the two-sided tail area

$$\Pr(U_0 \geq |u_0|) + \Pr(U_0 \leq -|u_0|).$$

For example, suppose that $n = 20$ and the calculated value of U_0 is 2.1. It is found from a table of the t distribution with 18 degrees of freedom that the corresponding tail area is 0.05. Hence, at each level of significance $\alpha_0 < 0.05$, the null hypothesis H_0 would not be rejected. At every level of significance $\alpha_0 \geq 0.05$, H_0 would be rejected.

Tests of Hypotheses about β_1 Let β_1^* be a specified number ($-\infty < \beta_1^* < \infty$), and suppose that it is desired to test the following hypotheses about the regression

coefficient β_1:

$$H_0: \quad \beta_1 = \beta_1^*,$$
$$H_1: \quad \beta_1 \neq \beta_1^*. \tag{11.3.21}$$

These hypotheses are the same as those in Eq. (11.3.13) if we make the substitutions $c_0 = 0$, $c_1 = 1$, and $c_* = \beta_1^*$. If we substitute these values into the formula for U_{01} in Eq. (11.3.14), we obtain the following random variable, U_1,

$$U_1 = s_x \frac{\hat{\beta}_1 - \beta_1^*}{\sigma'}, \tag{11.3.22}$$

which then has the t distribution with $n - 2$ degrees of freedom if H_0 is true.

Suppose that in a problem of simple linear regression we are interested in testing the hypothesis that the variable Y is actually unrelated to the variable X. Under Assumptions 11.2.1–11.2.5, this hypothesis is equivalent to the hypothesis that the regression function $E(Y|x)$ is constant and not actually a function of x. Since it is assumed that the regression function has the form $E(Y|x) = \beta_0 + \beta_1 x$, this hypothesis is in turn equivalent to the hypothesis that $\beta_1 = 0$. Thus, the problem is one of testing the following hypotheses:

$$H_0: \quad \beta_1 = 0,$$
$$H_1: \quad \beta_1 \neq 0.$$

Here the hypothesized value β_1^* is 0.

Let u_1 denote the value of U_1 calculated from a given set of observed values (x_i, y_i) for $i = 1, \ldots, n$. Then the p-value corresponding to these data is

$$\Pr(U_1 \geq |u_1|) + \Pr(U_1 \leq -|u_1|).$$

Example 11.3.2

Gasoline Mileage. Consider the two variables gasoline mileage and engine horsepower in Example 11.1.4. This time, let Y be 1 over gasoline mileage, that is, gallons per mile. Also, let X be engine horsepower. A plot of the observed (x_i, y_i) pairs is given in Fig. 11.8 together with the fitted least-squares regression line. Notice how much straighter the relationship is between the two variables in Fig. 11.8 than between the two variables in Fig. 11.6. The least-squares estimates for a simple linear regression of gallons per mile on engine horsepower are $\hat{\beta}_0 = 0.01537$ and $\hat{\beta}_1 = 1.396 \times 10^{-4}$. Also, $\sigma' = 7.181 \times 10^{-3}$, $\overline{x} = 183.97$, and $s_x = 1036.9$. Suppose that

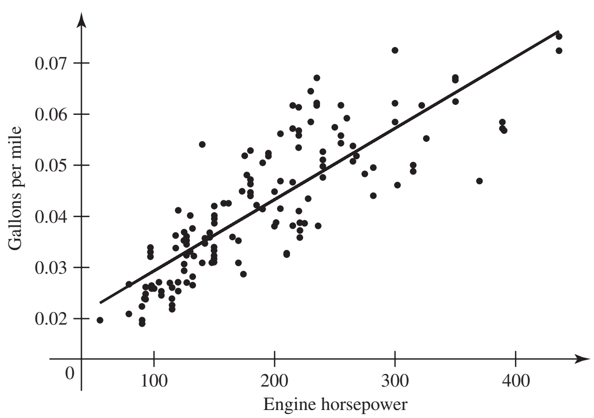


Figure 11.8 Plot of gallons per mile versus engine horsepower for 173 cars in Example 11.3.2. The least-squares regression line is drawn on the plot.

we wanted to test the null hypothesis $H_0 : \beta_1 \ge 0$ against the alternative $H_1 : \beta_1 < 0$. The observed value of the statistic U_1 in Eq. (11.3.22) is

$$u_1 = 1036.9 \frac{1.396 \times 10^{-4} - 0}{7.139 \times 10^{-3}} = 20.15,$$

which is larger than the $1 - 10^{-16}$ quantile of the t distribution with 171 degrees of freedom. So, we would reject H_0 at every level $\alpha_0 \le 10^{-16}$. ◀

Tests of Hypotheses about the Mean of a Future Observation Suppose that we are interested in testing the hypothesis that the regression line $y = \beta_0 + \beta_1 x$ passes through a particular point (x^*, y^*), where $x^* \ne 0$. In other words, suppose that we are interested in testing the following hypotheses:

$$\begin{aligned} H_0 &: \quad \beta_0 + \beta_1 x^* = y^*, \\ H_1 &: \quad \beta_0 + \beta_1 x^* \ne y^*. \end{aligned}$$

These hypotheses have the same form as the hypotheses (11.3.13) with $c_0 = 1$, $c_1 = x^*$, and $c_* = y^*$. Hence, they can be tested by carrying out a t test with $n - 2$ degrees of freedom that is based on the statistic U_{01}.

Example 11.3.3

Pressure and the Boiling Point of Water. In Example 11.3.1, the traveler was interested in testing the null hypothesis that $H_0 : \beta_0 + 201.5\beta_1 = 24.5$ versus $H_1 : \beta_0 + 201.5\beta_1 \ne 24.5$. We shall make use of the statistic U_{01} in Eq. (11.3.14) with $c_0 = 1$ and $c_1 = 201.5$. Based on the data in Table 11.5, we have already computed the least-squares estimates $\hat{\beta}_0 = -81.049$ and $\hat{\beta}_1 = 0.5228$. We can also compute $n = 17$, $s_x^2 = 530.78$, $\overline{x} = 202.95$, and $\sigma' = 0.2328$. Then

$$U_{01} = \left[\frac{1}{17} + \frac{(202.95 - 201.5)^2}{530.78} \right]^{1/2} \frac{-81.049 + 201.5 \times 0.5228 - 24.5}{0.2328} = -0.2204.$$

If H_0 is true, then $U_{0,1}$ has the t distribution with $n - 2 = 15$ degrees of freedom. The p-value corresponding to the observed value -0.2204 is 0.8285. The null hypothesis would be rejected at level α_0 only if $\alpha_0 \ge 0.8285$. ◀

Confidence Intervals

A confidence interval for β_0, β_1, or any linear combination of the two can be obtained from the corresponding test procedure.

Theorem 11.3.5

Let c_0 and c_1 be scalar constants that are not both 0. The open interval between the two random variables

$$c_0\hat{\beta}_0 + c_1\hat{\beta}_1 \pm \sigma' \left[\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2} \right]^{1/2} T_{n-2}^{-1}\left(1 - \frac{\alpha_0}{2}\right) \qquad (11.3.23)$$

is a coeficient $1 - \alpha_0$ confidence interval for $c_0\beta_0 + c_1\beta_1$.

Proof Consider the general hypotheses (11.3.13). Theorem 9.1.1 tells us that the set of all values of c_* for which the null hypothesis H_0 would not be rejected at the level of significance α_0 forms a confidence interval for $c_0\beta_0 + c_1\beta_1$ with confidence coefficient $1 - \alpha_0$. It is straightforward to check that c_* is between the two random variables in (11.3.23) if and only if $|U_{01}| < T_{n-2}^{-1}(1 - \alpha_0/2)$, which specifies when the level α_0 would not reject H_0 according to Theorem 11.3.3. ■

Example 11.3.4

Gasoline Mileage. In Example 11.3.2, we rejected the null hypothesis that $\beta_1 \le 0$, but we might wish to form an interval estimate of β_1. Apply Theorem 11.3.5 with $c_0 = 0$ and $c_1 = 1$. The endpoints of a coefficient $1 - \alpha_0$ confidence interval are then

$$\hat{\beta}_1 \pm \frac{\sigma'}{s_x} T^{-1}_{n-2}\left(1 - \frac{\alpha_0}{2}\right).$$

For example, suppose that we desire a coefficient 0.8 confidence interval for β_1. We find $T^{-1}_{171}(0.9) = 1.287$ using computer software (or we could have interpolated in the table in the back of the text). The remaining values needed to compute the endpoints are given in Example 11.3.2, and the observed interval is $(1.307 \times 10^{-4}, 1.485 \times 10^{-4})$. ◀

Other special cases of Theorem 11.3.5 are when $c_0 = 1$ and $c_1 = 0$, which provides a confidence interval for β_0, and when $c_0 = 1$ and $c_1 = x$, which provides a confidence interval for the mean of Y when $X = x$. The second of these can also be described as the height $\theta = \beta_0 + \beta_1 x$ of the regression line at a given point x. The corresponding confidence interval has the endpoints

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm T^{-1}_{n-2}\left(1 - \frac{\alpha_0}{2}\right) \sigma' \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{s_x^2}\right]^{1/2}. \tag{11.3.24}$$

Prediction Intervals On page 703, we discussed predicting a new Y value (independent of the observed data) when we knew the corresponding value of x. Suppose that we want an interval that should contain Y with some specified probability $1 - \alpha_0$. We can construct such an interval by considering the joint distribution of Y, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$, and S^2.

Theorem 11.3.6

In the simple linear regression problem, let Y be a new observation with predictor x such that Y is independent of $Y_1, \ldots, Y_n$. Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Then the probability that Y is between the following two random variables is $1 - \alpha_0$:

$$\hat{Y} \pm T^{-1}_{n-2}\left(1 - \frac{\alpha_0}{2}\right) \sigma' \left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{s_x^2}\right]^{1/2}. \tag{11.3.25}$$

Proof Since Y is independent of the observed data, we have that Y, $\hat{Y}$, and S^2 are all independent. Hence, the following two random variables are independent:

$$Z = \frac{Y - \hat{Y}}{\sigma\left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{s_x^2}\right]^{1/2}}, \quad W = \frac{S^2}{\sigma^2}.$$

Since Y and $\hat{Y}$ are independent and normally distributed, Z has a normal distribution. Since $E(Y) = E(\hat{Y})$, the mean of Z is 0. It follows from Eq. (11.2.13) that the variance of Z is 1. It follows from Theorem 11.3.2 that W has the χ^2 distribution with $n - 2$ degrees of freedom. It follows that $Z/(W/[n-2])^{1/2}$ has the t distribution with $n - 2$ degrees of freedom. It is easy to see that $Z/(W/[n-2])^{1/2}$ is the same as

$$U_x = \frac{Y - \hat{Y}}{\sigma'\left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{s_x^2}\right]^{1/2}}. \tag{11.3.26}$$

It follows that $\Pr(|U_x| < T^{-1}_{n-2}(1-\alpha_0/2)) = 1-\alpha_0$. It is then straightforward to show that Y is between the two random variables in (11.3.25) if and only if $|U_x| < T^{-1}_{n-2}(1-\alpha_0/2)$. ■

Definition 11.3.1 Prediction Interval. The random interval whose endpoints are given by (11.3.25) is called a coefficient $1-\alpha_0$ *prediction interval* for Y.

Prior to observing the data, when σ', $\hat{\beta}_0$, $\hat{\beta}_1$, and Y are all still random variables, the endpoints in (11.3.25) have the property that the probability is $1-\alpha_0$ that Y will be between the endpoints, and hence in the interval. After the data are observed, the interpretation of the interval whose endpoints are in (11.3.25) is similar to the interpretation of a confidence interval, but with the added complication that Y is still a random variable.

Example 11.3.5 Gasoline Mileage. Suppose that we wish to predict the gasoline mileage for a car with a particular engine horsepower x in Example 11.3.2. In particular, let $x = 100$, and we shall use $\alpha_0 = 0.1$ to form a prediction interval as above. Using the values computed in Example 11.3.2 and Eq. (11.3.25), we obtain the interval (0.01737, 0.04127) for predicting Y gallons per mile. Since Y is in this interval if and only if $1/Y$ is between $1/0.01737 = 57.56$ and $1/0.04127 = 24.23$, we can claim that the following interval is the observed value of a 90 percent prediction interval for miles per gallon: (24.23, 57.56). ◄

The Analysis of Residuals

Whenever a statistical analysis is carried out, it is important to verify that the observed data appear to satisfy the assumptions on which the analysis is based. For example, in the statistical analysis of a problem of simple linear regression, we have assumed that the regression of Y on X is a linear function and that the observations $Y_1, \dots, Y_n$ are independent. The M.L.E.'s of β_0 and β_1 and the tests of hypotheses about β_0 and β_1 were developed on the basis of these assumptions, but the data were not examined to find out whether or not these assumptions were reasonable.

One way to make a quick and informal check of these assumptions is to examine the discrepancies between the observed values $y_1, \dots, y_n$ and the fitted regression line.

Definition 11.3.2 Residuals/Fitted Values. For $i = 1, \dots, n$, the observed values of $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ are called the *fitted values*. For $i = 1, \dots, n$, the observed values of $e_i = y_i - \hat{y}_i$ are called the *residuals*.

Specifically, suppose that the n points (x_i, e_i), for $i = 1, \dots, n$ are plotted in the xe-plane. It must be true (see Exercise 4 at the end of Sec. 11.1) that $\sum_{i=1}^n e_i = 0$ and $\sum_{i=1}^n x_i e_i = 0$. However, subject to these restrictions, the positive and negative residuals should be scattered randomly among the points (x_i, e_i). If the positive residuals e_i tend to be concentrated at either the extreme values of x_i or the central values of x_i, then either the assumption that the regression of Y on X is a linear function or the assumption that the observations $Y_1, \dots, Y_n$ are independent may be violated. In fact, if the plot of the points (x_i, e_i) exhibits any type of regular pattern, the assumptions may be violated.

Example 11.3.6 Pressure and the Boiling Point of Water. The residuals from a least-squares fit to the data in Example 11.2.2 can be computed using the coefficients reported in Example 11.2.5: $\hat{\beta}_0 = -81.06$ and $\hat{\beta}_1 = 0.5229$. Table 11.7 contains the original data together

Table 11.7 Data from Table 11.5 together with fitted values, residuals from least-squares fit, and logarithm of pressure

x_i	y_i	$\hat{y}_i = -81.06 + 0.5229x_i$	$e_i = y_i - \hat{y}_i$	$\log(y_i)$
194.5	20.79	20.64	0.1512	3.034
194.3	20.79	20.53	0.2557	3.034
197.9	22.40	22.42	−0.0167	3.109
198.4	22.67	22.68	−0.0081	3.121
199.4	23.15	23.20	−0.0510	3.142
199.9	23.35	23.46	−0.1125	3.151
200.9	23.89	23.99	−0.0954	3.173
201.1	23.99	24.09	−0.0999	3.178
201.4	24.02	24.25	−0.2268	3.179
201.3	24.01	24.19	−0.1845	3.178
203.6	25.14	25.40	−0.2572	3.224
204.6	26.57	25.92	0.6499	3.280
209.5	28.49	28.48	0.0078	3.350
208.6	27.76	28.01	−0.2516	3.324
210.7	29.04	29.11	−0.0697	3.369
211.9	29.88	29.74	0.1428	3.397
212.2	30.06	29.89	0.1660	3.403

with the fitted values $\hat{y}_i = -81.06 + 0.5229x_i$ and the residuals $e_i = y_i - \hat{y}_i$ for all i. A plot of the residuals versus boiling point is shown in Fig. 11.9. This plot has two striking features. One is the exceptionally large positive residual corresponding to $x_i = 204.6$ at the top of the plot. Observations with such large residuals are sometimes called *outliers*. Perhaps either the x_i or y_i value corresponding to this observation was recorded incorrectly or this observation was taken under conditions different from those of the other observations. Or perhaps that particular y_i value just happened to be very far from its mean. The other striking feature of the plot is that, aside from the outlier, the other residuals seem to form a U-shaped pattern. This sort of pattern suggests that the relationship between the two variables might be better described by a curve rather than a straight line.

Techniques for dealing with the two features that we noticed in Fig. 11.9 can be found in books devoted to regression methodology such as Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1982), Draper and Smith (1998), and Weisberg (1985). One possible technique to deal with the curved look of the residual plot is to transform one or both of the two variables Y and X before performing the regression. Indeed, Forbes (1857) suspected that the logarithm of pressure would be linearly related to boiling point. Table 11.7 also contains the logarithms of pressure. If we perform a regression of the logarithm of pressure on the boiling point, we obtain the least-squares estimates $\hat{\beta}_0 = -0.9709$ and $\hat{\beta}_1 = 0.0206$. The observed value of σ' is 8.730×10^{-3}. Residuals from this fit can be computed as $\log(y_i) - (-0.9709 + 0.0206x_i)$, and they are plotted in Fig. 11.10. The one large residual still appears in

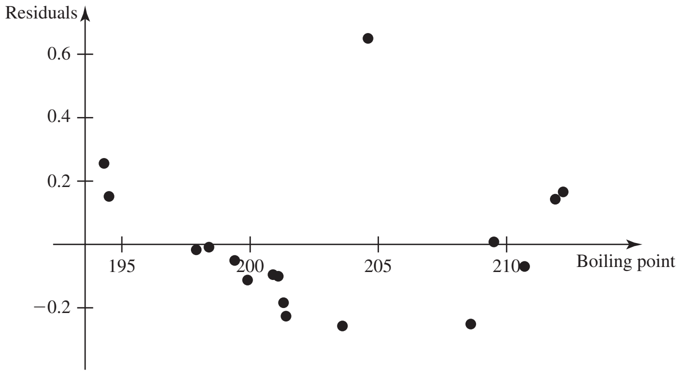


Figure 11.9 Plot of residuals versus boiling point for Example 11.3.6.

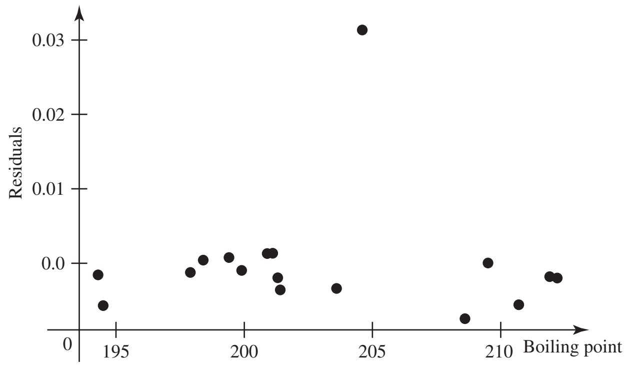


Figure 11.10 Plot of residuals from regression of log-pressure versus boiling point for Example 11.3.6.

Fig. 11.10, but the curved shape of the remaining residuals has vanished. To see what effect that one observation has on the regression, we can fit the regression using only the other 16 observations. In this case, the estimated coefficients are $\hat{\beta}_0 = -0.9518$ and $\hat{\beta}_1 = 0.0205$ with $\sigma' = 2.616 \times 10^{-3}$. The coefficients don't change much, but the estimated standard deviation drops to less than one-third of its previous value. ◀

Note: Both Models Cannot Be Correct in Example 11.3.6. It cannot be the case that both the mean of pressure and the mean of the logarithm of pressure are linear functions of boiling point. When the residual plot in Fig. 11.9 revealed a curved shape, we began to suspect that the mean of pressure was *not* a linear function of boiling point. In this case, the probabilistic calculations performed in Examples 11.2.2, 11.2.5, and 11.3.3 become suspect as well.

Note: What to Do with Outliers. The data point with $X = 204.6$ in Example 11.3.6 makes it difficult to interpret the results of the regression analysis. Forbes (1857) labels this point "Evidently a mistake." Generally, when such data points appear in our data sets, we should try to verify whether they were collected under the same conditions as the remaining data. Sometimes the process by which the data are collected changes during the experiment. If the removal of the outlier makes a noticeable difference to the analysis, then that observation must be dealt with. If it is not possible to show that the observation should be removed based on how it was collected, it might be that the distribution of the Y_i values is different from a normal distribution. It might be that the distribution has higher probability of producing

extremely large deviations from the mean. In this case, one might have to resort to robust regression procedures similar to the robust procedures described in Sec. 10.7. Interested readers should consult Hampel et al. (1986) or Rousseeuw and Leroy (1987).

Normal Quantile Plots Another plot that is helpful in assessing the assumptions of the regression model is the *normal quantile plot*, sometimes called a normal scores plot or a normal Q-Q plot. Assume that the residuals are reasonable estimates of $\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$. Each ε_i has the normal distribution with mean 0 and variance σ^2 according to the linear regression model. The normal quantile plot compares quantiles of a normal distribution with the ordered values of the residuals. We expect about 25 percent of the residuals to be below the 0.25 quantile of the normal distribution. We expect about 80 percent of the residuals to be below the 0.8 quantile of the normal distribution, and so forth. We can see how closely these expectations are met by plotting the ordered residuals against quantiles of the normal distribution.

Let $r_1 \le r_2 \le \cdots \le r_n$ be the residuals ordered from smallest to largest. The points that we plot are $(\Phi^{-1}(i/[n+1]), r_i)$ for $i = 1, \ldots, n$, where Φ^{-1} is the standard normal quantile function. The numbers $\Phi^{-1}(i/[n+1])$ for $i = 1, \ldots, n$ are n quantiles of the standard normal distribution that divide the standard normal distribution into intervals of equal probability, including the intervals below the first quantile and above the last one. If the plotted points lie roughly along the line $y = x$, then roughly 25 percent of the residuals lie below the 0.25 quantile of the standard normal distribution, and roughly 80 percent of the residuals lie below the 0.8 quantile, and so on. If the points lie on a different line $y = ax + b$, then we could multiply the first coordinate of each point by a and add b to the first coordinate. This would make the new points lie on the line $y = x$, and the first coordinate of each point is now a quantile of the normal distribution with mean b and variance a^2. So roughly 25 percent of the residuals lie below the 0.25 quantile of the normal distribution with mean b and variance a^2, and so on. So, we examine the normal quantile plot to see how close the points are to lying on a straight line. We don't care which line it is, because we only care whether the data look like they come from some normal distribution. We fit the regression model to help decide which normal distribution.

Example 11.3.7

Pressure and the Boiling Point of Water. As an illustration of the normal quantile plot, we deleted the troublesome observation (number 12) from the data set of Example 11.3.6 and fit the model in which the logarithm of pressure is regressed on the boiling point. The resulting normal quantile plot is shown in Fig. 11.11. The points in Fig. 11.11 lie roughly on a line, although it is not difficult to detect some curvature in the plot. It is usually the case that the extreme residuals (lowest and highest) do not line up well with the others, so one normally pays closest attention to the middle of the plot. Extreme observations that fall very far from the pattern of the others suggest a more serious problem. Outliers will typically show up in this way as well as in the other residual plots. ◀

If we know the order in which the observations were taken, there are some additional plots that can help reveal whether there is some dependence between the observations. We will introduce these plots when we discuss multiple regression later in this chapter. Readers desiring a deeper understanding of graphics associated with linear regression should read Cook and Weisberg (1994).

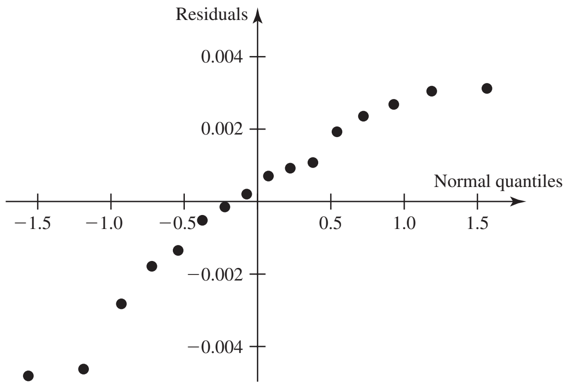


Figure 11.11 Normal quantile plot for regression of log-pressure on boiling point with observation number 12 removed.

Inference about Both β_0 and β_1 Simultaneously

Tests of Hypotheses about Both β_0 and β_1 Suppose next that β_0^* and β_1^* are given numbers and that we are interested in testing the following hypotheses about the values of β_0 and β_1:

$$\begin{aligned} H_0:&\quad \beta_0 = \beta_0^* \quad \text{and} \quad \beta_1 = \beta_1^*, \\ H_1:&\quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{11.3.27}$$

These hypotheses are not a special case of (11.3.13); hence, we shall not be able to test these hypotheses using U_{01} from Eq. (11.3.14). Instead, we shall derive the likelihood ratio test procedure for the hypotheses (11.3.27).

The likelihood function $f_n(\boldsymbol{y}|\boldsymbol{x}, \beta_0, \beta_1, \sigma^2)$ is given by Eq. (11.2.2). We know from Sec. 11.2 that the likelihood function attains its maximum value when β_0, β_1, and σ^2 are equal to the M.L.E.'s $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$, as given by Eq. (11.1.1) and Eq. (11.2.3).

When the null hypothesis H_0 is true, the values of β_0 and β_1 must be β_0^* and β_1^*, respectively. For these values of β_0 and β_1, the maximum value of $f_n(\boldsymbol{y}|\boldsymbol{x}, \beta_0^*, \beta_1^*, \sigma^2)$ over all the possible values of σ^2 will be attained when σ^2 has the following value $\hat{\sigma}_0^2$:

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0^* - \beta_1^* x_i)^2.$$

Now consider the statistic

$$\Lambda(\boldsymbol{y}|\boldsymbol{x}) = \frac{\sup_{\sigma^2} f_n(\boldsymbol{y}|\boldsymbol{x}, \beta_0^*, \beta_1^*, \sigma^2)}{\sup_{\beta_0,\beta_1,\sigma^2} f_n(\boldsymbol{y}|\boldsymbol{x}, \beta_0, \beta_1, \sigma^2)}.$$

By using the results that have just been described, it can be shown that

$$\Lambda(\boldsymbol{y}|\boldsymbol{x}) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} = \left[\frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^{n}(y_i - \beta_0^* - \beta_1^* x_i)^2}\right]^{n/2}. \tag{11.3.28}$$

The denominator of the final expression in Eq. (11.3.28) can be rewritten as follows:

$$\begin{aligned}&\sum_{i=1}^{n}(y_i - \beta_0^* - \beta_1^* x_i)^2 \\ &\quad = \sum_{i=1}^{n}[(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) + (\hat{\beta}_0 - \hat{\beta}_0^*) + (\hat{\beta}_1 - \hat{\beta}_1^*)x_i]^2.\end{aligned} \tag{11.3.29}$$

To simplify this expression further, let the statistic S^2 be defined by Eq. (11.3.9), and let the statistic Q^2 be defined as follows:

$$\begin{aligned}Q^2 = n(\hat{\beta}_0 - \beta_0^*)^2 + \left(\sum_{i=1}^{n} x_i^2\right)(\hat{\beta}_1 - \beta_{\beta_1}^*)^2 \\ + 2n\overline{x}(\hat{\beta}_0 - \beta_0^*)(\hat{\beta}_1 - \beta_1^*).\end{aligned} \tag{11.3.30}$$

We shall now expand the right side of Eq. (11.3.29) and use the following relations, which were established in Exercise 4 of Sec. 11.1:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

We then obtain the relation

$$\sum_{i=1}^{n}(y_i - \beta_0^* - \beta_1^* x_i)^2 = S^2 + Q^2.$$

It now follows from Eq. (11.3.28) that

$$\Lambda(\boldsymbol{y}|\boldsymbol{x}) = \left(\frac{S^2}{S^2 + Q^2}\right)^{n/2} = \left(1 + \frac{Q^2}{S^2}\right)^{-n/2}. \tag{11.3.31}$$

The likelihood ratio test procedure specifies rejecting H_0 when $\Lambda(\boldsymbol{y}|\boldsymbol{x}) \le k$. It can be seen from Eq. (11.3.31) that this procedure is equivalent to rejecting H_0 when $Q^2/S^2 \ge k'$, where k' is a suitable constant. To put this procedure in a more standard form, we shall let the statistic U^2 be defined as follows:

$$U^2 = \frac{\frac{1}{2}Q^2}{\sigma'^2}. \tag{11.3.32}$$

Then the likelihood ratio test procedure specifies rejecting H_0 when $U^2 \ge \gamma$, where γ is a suitable constant.

We shall now determine the distribution of the statistic U^2 when the hypothesis H_0 is true. It can be shown (see Exercises 7 and 8) that when H_0 is true, the random variable Q^2/σ^2 has the χ^2 distribution with two degrees of freedom. Also, because the random variable S^2 and the random vector $(\hat{\beta}_0, \hat{\beta}_1)$ are independent, and because Q^2 is a function of $\hat{\beta}_0$ and $\hat{\beta}_1$, it follows that the random variables Q^2 and S^2 are independent. Finally, we know that S^2/σ^2 has the χ^2 distribution with $n-2$ degrees of freedom. Therefore, when H_0 is true, the statistic U^2 defined by Eq. (11.3.32) will have the F distribution with 2 and $n-2$ degrees of freedom. Since the null hypothesis H_0 is rejected if $U^2 \ge \gamma$, the value of γ corresponding to a specified level of significance α_0 $(0 < \alpha_0 < 1)$ will be the $1-\alpha_0$ quantile of this F distribution, namely, $F_{2,n-2}^{-1}(1-\alpha_0)$.

Joint Confidence Set Next, consider the problem of constructing a confidence set for the pair of unknown regression coefficients β_0 and β_1. Such a confidence set can

be obtained from the statistic U^2 defined by Eq. (11.3.32), which was used to test the hypotheses (11.3.27). Specifically, let $F^{-1}_{2,n-2}(1-\alpha_0)$ be the $1-\alpha_0$ quantile of the F distribution with 2 and $n-2$ degrees of freedom. Then the set of all pairs of values of β_0^* and β_1^* such that $U^2 < F^{-1}_{2,n-2}(1-\alpha_0)$ will form a confidence set for the pair (β_0, β_1) with confidence coefficient $1-\alpha_0$. It can be shown (see Exercise 16) that this confidence set will contain all the points (β_0, β_1) inside a certain ellipse in the $\beta_0\beta_1$-plane. In other words, this confidence set will actually be a confidence ellipse.

The confidence ellipse that has just been derived for β_0 and β_1 can be used to construct a confidence set for the entire regression line $y = \beta_0 + \beta_1 x$. Corresponding to each point (β_0, β_1) inside the ellipse, we can draw a straight line $y = \beta_0 + \beta_1 x$ in the xy-plane. The collection of all these straight lines corresponding to all points (β_0, β_1) inside the ellipse will be a confidence set with confidence coefficient $1-\alpha_0$ for the actual regression line. A rather lengthy and detailed analysis, which will not be presented here [see Scheffé (1959, section 3.5)], shows that the upper and lower limits of this confidence set are the curves defined by the following relations:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \pm [2F^{-1}_{2,n-2}(1-\alpha_0)]^{1/2}\sigma' \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{s_x^2}\right]^{1/2}. \tag{11.3.33}$$

In other words, with confidence coefficient $1-\alpha_0$, the actual regression line $y = \beta_0 + \beta_1 x$ will lie between the curve obtained by using the plus sign in (11.3.33) and the curve obtained by using the minus sign in (11.3.33). The region between these curves is often called a *confidence band* or *confidence belt* for the regression line.

In similar fashion, the confidence ellipse can be used to construct simultaneous confidence intervals for every linear combination of β_0 and β_1. The coefficient $1-\alpha_0$ interval for $c_0\beta_0 + c_1\beta_1$ has the endpoints

$$c_0\hat{\beta}_0 + c_1\hat{\beta}_1 \pm \sigma' \left[\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2}\right]^{1/2} \left[2F^{-1}_{2,n-2}(1-\alpha_0)\right]^{1/2}. \tag{11.3.34}$$

This differs from the individual confidence interval given in Eq. (11.3.23) solely in the replacement of the $1-\alpha_0/2$ quantile of the t_{n-2} distribution by the square root of 2 times the $1-\alpha_0$ quantile of the $F_{2,n-2}$ distribution. The simultaneous intervals are wider than the individual intervals because they satisfy a more restrictive requirement. The probability (prior to observing the data) is $1-\alpha_0$ that all of the intervals of the form (11.3.34) simultaneously contain their corresponding parameters. Each interval of the form (11.3.23) contains its corresponding parameter with probability $1-\alpha_0$, but the probability that two or more of them simultaneously contain their corresponding parameters is less than $1-\alpha_0$.

Alternative Tests and Confidence Sets The hypotheses (11.3.27) are a special case of (9.1.26), and they can be tested by the same method outlined immediately after (9.1.26). The resulting test leads to an alternative confidence set for the pair (β_0, β_1). The alternative level α_0 test of (11.3.27) merely combines the two level $\alpha_0/2$ tests of (11.3.20) and (11.3.21). To be specific, the alternative level α_0 test δ of (11.3.27) is to reject H_0 if either

$$|U_0| \ge T^{-1}_{n-2}\left(1 - \frac{\alpha_0}{4}\right) \text{ or } |U_1| \ge T^{-1}_{n-2}\left(1 - \frac{\alpha_0}{4}\right) \text{ or both,} \tag{11.3.35}$$

where U_0 and U_1 are, respectively, the statistics in (11.3.19) and (11.3.22) that would be used for testing (11.3.20) and (11.3.21).

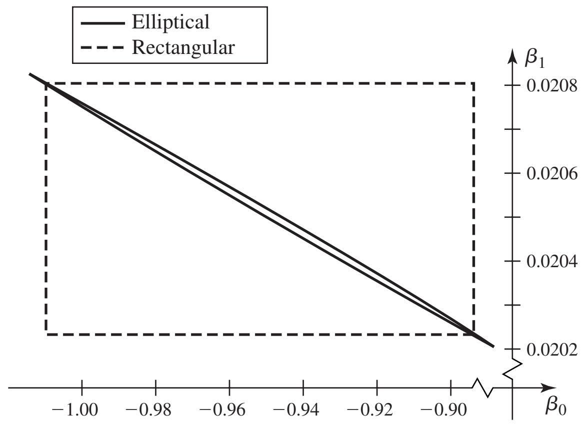


Figure 11.12 Elliptical and rectangular joint coefficient 0.95 confidence sets for (β_0, β_1) in Example 11.3.8.

The corresponding joint confidence set for (β_0, β_1) is the set of all (β_0^*, β_1^*) pairs such that both $|U_0|$ and $|U_1|$ are strictly less than $T_{n-2}^{-1}(1-\alpha_0/4)$. This alternative confidence set will be rectangular in shape rather than elliptical. This confidence rectangle also provides simulaneous confidence intervals for all linear combinations of the form $c_0\beta_0 + c_1\beta_1$. The formulas for the endpoints are not so pretty as (11.3.34). Let C be the joint confidence rectangle. Then the confidence interval for $c_0\beta_0 + c_1\beta_1$ is the following:

$$\left(\inf_{(\beta_0^*, \beta_1^*)\in C} c_0\beta_0^* + c_1\beta_1^*, \sup_{(\beta_0^*, \beta_1^*)\in C} c_0\beta_0^* + c_1\beta_1^* \right). \tag{11.3.36}$$

The sup and inf will each occur at one of the four corners of the rectangle, so one need only compute four values of $c_0\beta_0^* + c_1\beta_1^*$ to determine the interval. Some special cases are worked out in Exercise 24.

Example 11.3.8 Pressure and the Boiling Point of Water. In Examples 11.2.1 and 11.2.2, we computed the least-squares estimates and the variances and covariance of the estimates. Figure 11.12 shows both the elliptical and the rectangular coefficient 0.95 joint confidence sets for the pair (β_0, β_1). If all that we wanted were confidence intervals for the two parameters, we could extract those from both confidence sets. For the elliptical region, (11.3.34) gives the intervals $(-1.0149, -0.8886)$ and $(0.020207, 0.020830)$ for β_0 and β_1, respectively. Notice that the endpoints of these intervals are, respectively, the minimum and maximum values of β_0 and β_1 in the elliptical joint confidence set in Fig. 11.12. Similarly, the joint confidence intervals from the rectangular joint confidence set are, respectively, $(-1.0097, -0.8938)$ and $(0.020233, 0.020804)$, whose endpoints are also the minimum and maximum values of β_0 and β_1 in the rectangular joint confidence set in Fig. 11.12.

Finally, suppose that, in addition to confidence intervals for the two parameters β_0 and β_1, we also want a confidence band for the regression function, namely, the mean log-pressure at all temperatures x. This mean is of the form $c_0\beta_0 + c_1\beta_1$ with $c_0 = 1$ and $c_1 = x$. The confidence bands are plotted in Fig. 11.13 based both on the elliptical and rectangular joint confidence sets. For example, at $x = 201.5$, we get the intervals $(3.1809, 3.1846)$ and $(3.0672, 3.2983)$ from the elliptical and rectangular sets, respectively.

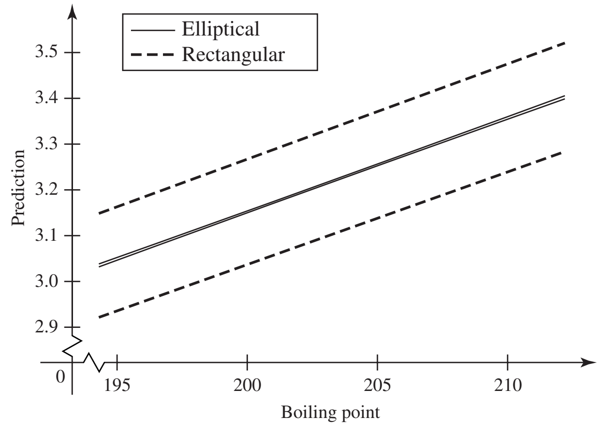


Figure 11.13 Coefficient 0.95 confidence bands for the regression function in Example 11.3.8. Bands are computed based both on the elliptical and on the rectangular joint confidence sets.

The joint confidence intervals for the two individual parameters are slightly shorter when computed from the rectangular confidence set compared to the elliptical set. But the confidence band for the regression function (Fig. 11.13) is much wider when computed from the rectangular set compared to the elliptical set. ◀

In Example 11.3.8, if one were interested solely in simultaneous confidence intervals for the three parameters β_0, β_1, and $\beta_0 + 201.5\beta_1$, instead of the entire regression function, one could obtain shorter intervals from a generalization of the rectangular joint confidence set. The generalization is based on the Bonferroni inequality from Theorem 1.5.8.

Theorem 11.3.7 Suppose that we are interested in forming simultaneous confidence intervals for several parameters $\theta_1, \ldots, \theta_n$. For each i, let (A_i, B_i) be a coefficient $1 - \alpha_i$ confidence interval for θ_i. Then the probability that all n confidence intervals simultaneously cover their corresponding parameters is at least $1 - \sum_{i=1}^{n} \alpha_i$.

Proof For each $i = 1, \ldots, n$, define the event $E_i = \{A_i < \theta_i < B_i\}$. Because (A_i, B_i) is a coefficient $1 - \alpha_i$ confidence interval for θ_i, we have $\Pr(E_i^c) \le \alpha_i$ for every i, and the probability that all n intervals simultaneously cover their corresponding parameters is $\Pr\left(\bigcap_{i=1}^{n} E_i\right)$. By the Bonferroni inequality, this last probability is at least $1 - \sum_{i=1}^{n} \alpha_i$. ■

Theorem 9.1.5 gives the corresponding result for a test of the joint hypotheses

$$H_0 : \theta_i = \theta_i^* \text{ for all } i\ ,\ H_1 : \text{ not } H_0, \tag{11.3.37}$$

If we want simultaneous coefficient $1 - \alpha_0$ confidence intervals for three parameters, let $\alpha_i = \alpha_0/3$.

Example 11.3.9 Pressure and the Boiling Point of Water. Suppose that we are interested solely in simultaneous coefficient 0.95 confidence intervals for the three parameters β_0, β_1, and $\beta_0 + 201.5\beta_1$ in Example 11.3.8. Then we can use coefficient $1 - 0.05/3 = 0.9833$ confidence intervals for each parameter. The necessary quantile of the t distribution is $T_{14}^{-1}(0.9917) = 2.7178$. The three intervals for β_0, β_1, and $\beta_0 + 201.5\beta_1$ are

(−1.0146, −0.8889), (0.020296, 0.020828), and (3.1809, 3.1845), respectively. Notice that these are all shorter than the corresponding intervals based on the elliptical joint confidence set. The first two of these intervals are longer than the corresponding intervals from the rectangular joint confidence set in Example 11.3.8, but the third interval is much shorter than the corresponding interval based on that same rectangular set. ◄

Finally, there is a way to construct a narrower confidence band for the entire regression function based on the Bonferroni inequality, but we leave the details to Exercise 25.

So, which confidence intervals should one use? Also, which test of (11.3.27) should one use? None of the tests that we have constructed are uniformly most powerful. Some are more powerful at some alternatives, while others are more powerful at other alternatives. The test corresponding to the rectangular joint confidence set is more powerful than the elliptical test if either β_0 or β_1 is a little larger or smaller than its hypothesized value while the other parameter is close to its hypothesized value. The elliptical test is more powerful than the rectangular test if both β_0 and β_1 are a little different from their hypothesized values, even if neither is far enough away to cause the rectangular test to reject. Without any specification of which alternatives are most important to detect, one might choose the elliptical test. On the other hand, if one's sole need is for a few confidence intervals and not a confidence band for the entire regression function, the intevals based on the Bonferroni inequality will generally be shorter. The different tests and confidence intervals differ solely by which quantiles are used in their construction. The larger the quantile, the longer the confidence interval. Table 11.8 gives the quantiles needed for the intervals based on the elliptical joint confidence set (which do *not* depend on how many intervals one constructs) and the quantiles needed for various numbers of intervals based on the Bonferroni inequality. One can see that the Bonferroni intervals will generally be shorter if one wants only three or fewer.

Summary

For constants c_0 and c_1 that are not both 0, we saw that

$$\left[\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2}\right]^{-1/2} \frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - (c_0\beta_0 + c_1\beta_1)}{\sigma'} \tag{11.3.38}$$

has the t distribution with $n-2$ degrees of freedom under the assumptions of simple linear regression. We can use the random variable in (11.3.38) to test hypotheses about or to construct confidence intervals for β_0, β_1, and linear combinations of the two. We also learned how to form a prediction interval for a future observation Y when the corresponding value for X is known.

Tests about both β_0 and β_1 simultaneously are based on the statistic U^2 in Eq. (11.3.32), which has the F distribution with 2 and $n-2$ degrees of freedom when the null hypothesis H_0 in Eq. (11.3.27) is true. A confidence band for the entire regression line $y = \beta_0 + \beta_1 x$ (a collection of confidence intervals, one for each x, such that all of the intervals simultaneously cover the true values of $\beta_0 + \beta_1 x$ with probability $1 - \alpha_0$) is given by Eq. (11.3.33). The intervals in the confidence band are slightly wider than the individual confidence intervals with each separate x.

Table 11.8 Comparison of the quantiles needed to compute k simultaneous joint confidence intervals based on the Bonferroni inequality and based on the elliptical joint confidence set

		$T^{-1}_{n-2}(1-\alpha_0/[2k])$				
α_0	n	$k=1$	$k=2$	$k=3$	$k=4$	$[2F^{-1}_{2,n-2}(1-\alpha_0)]^{1/2}$
0.05	5	3.18	4.18	4.86	5.39	4.37
	10	2.31	2.75	3.02	3.21	2.99
	15	2.16	2.53	2.75	2.90	2.76
	20	2.10	2.45	2.64	2.77	2.67
	60	2.00	2.30	2.47	2.58	2.51
	120	1.98	2.27	2.43	2.54	2.48
	∞	1.96	2.24	2.40	2.50	2.45
0.01	5	5.84	7.45	8.58	9.46	7.85
	10	3.36	3.83	4.12	4.33	4.16
	15	3.01	3.37	3.58	3.73	3.66
	20	2.88	3.20	3.38	3.51	3.47
	60	2.66	2.92	3.06	3.16	3.16
	120	2.62	2.86	3.00	3.09	3.10
	∞	2.58	2.81	2.94	3.03	3.04

It is good practice to plot residuals from a regression against the predictor X. Such plots can reveal evidence of departures from the assumptions that underly the distribution theory developed in this section. In particular, one should look for patterns and unusual points in the plot of residuals. Plots of residuals against X help reveal departures from the assumed form of the mean of Y. Plots of sorted residuals against normal quantiles help reveal departures from the assumption that the distribution of each Y_i is normal.

Exercises

1. Suppose that in a problem of simple linear regression, the 10 pairs of observed values of x_i and y_i given in Table 11.9 are obtained. Test the following hypotheses at the level of significance 0.05:

$$H_0: \quad \beta_0 = 0.7,$$
$$H_1: \quad \beta_0 \neq 0.7.$$

2. For the data presented in Table 11.9, test at the level of significance 0.05 the hypothesis that the regression line passes through the origin in the xy-plane.

3. For the data presented in Table 11.9, test at the level of significance 0.05 the hypothesis that the slope of the regression line is 1.

Table 11.9 Data for Exercise 1

i	x_i	y_i	i	x_i	y_i
1	0.3	0.4	6	1.0	0.8
2	1.4	0.9	7	2.0	0.7
3	1.0	0.4	8	−1.0	−0.4
4	−0.3	−0.3	9	−0.7	−0.2
5	−0.2	0.3	10	0.7	0.7

4. For the data presented in Table 11.9, test at the level of significance 0.05 the hypothesis that the regression line is horizontal.

5. For the data presented in Table 11.9, test the following hypotheses at the level of significance 0.10:

$$H_0: \quad \beta_1 = 5\beta_0,$$
$$H_1: \quad \beta_1 \neq 5\beta_0.$$

6. For the data presented in Table 11.9, test the hypothesis that when $x = 1$, the height of the regression line is $y = 1$ at the level of significance 0.01.

7. In a problem of simple linear regression, let $D = \hat{\beta}_0 + \hat{\beta}_1\overline{x}$. Show that the random variables $\hat{\beta}_1$ and D are uncorrelated, and explain why $\hat{\beta}_1$ and D must therefore be independent.

8. Let the random variable D be defined as in Exercise 7, and let the random variable Q^2 be defined by Eq. (11.3.30).

a. Show that

$$\frac{Q^2}{\sigma^2} = \frac{(\hat{\beta}_1 - \beta_1^*)^2}{\text{Var}(\hat{\beta}_1)} + \frac{(D - \beta_0^* - \beta_1^*\overline{x})^2}{\text{Var}(D)}.$$

b. Explain why the random variable Q^2/σ^2 will have the χ^2 distribution with two degrees of freedom when the hypothesis H_0 in (11.3.27) is true.

9. For the data presented in Table 11.9, test the following hypotheses at the level of significance 0.05:

H_0: $\beta_0 = 0$ and $\beta_1 = 1$,

H_1: At least one of the values $\beta_0 = 0$ and $\beta_1 = 1$ is incorrect.

10. For the data presented in Table 11.9, construct a confidence interval for β_0 with confidence coefficient 0.95.

11. For the data presented in Table 11.9, construct a confidence interval for β_1 with confidence coefficient 0.95.

12. For the data presented in Table 11.9, construct a confidence interval for $5\beta_0 - \beta_1 + 4$ with confidence coefficient 0.90.

13. For the data presented in Table 11.9, construct a confidence interval with confidence coefficient 0.99 for the height of the regression line at the point $x = 1$.

14. For the data presented in Table 11.9, construct a confidence interval with confidence coefficient 0.99 for the height of the regression line at the point $x = 0.42$.

15. Suppose that in a problem of simple linear regression, a confidence interval with confidence coefficient $1 - \alpha_0$ $(0 < \alpha_0 < 1)$ is constructed for the height of the regression line at a given value of x. Show that the length of this confidence interval is shortest when $x = \overline{x}$.

16. Let the statistic U^2 be as defined by Eq. (11.3.32), and let γ be fixed positive constant. Show that for all observed values (x_i, y_i), for $i = 1, \dots, n$, the set of points (β_0^*, β_1^*) such that $U^2 < \gamma$ is the interior of an ellipse in the $\beta_0^*\beta_1^*$-plane.

17. For the data presented in Table 11.9, construct a confidence ellipse for β_0 and β_1 with confidence coefficient 0.95.

18.

a. For the data presented in Table 11.9, sketch a confidence band in the xy-plane for the regression line with confidence coefficient 0.95.

b. On the same graph, sketch the curves which specify the limits at each point x of a confidence interval with confidence coefficient 0.95 for the value of the regression line at the point x.

19. Determine a value of c such that in a problem of simple linear regression, the statistic $c\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_2 x_i)^2$ will be an unbiased estimator of σ^2.

20. Suppose that a simple linear regression of miles per gallon (Y) on car weight (X) has been performed with $n = 32$ observations. Suppose that the least-squares estimates are $\hat{\beta}_0 = 68.17$ and $\hat{\beta}_1 = -1.112$, with $\sigma' = 4.281$. Other useful statistics are $\overline{x} = 30.91$, and $\sum_{i=1}^{n}(x_i - \overline{x})^2 = 2054.8$.

a. Suppose that we want to predict miles per gallon Y for a new observation with weight $X = 24$. What would be our prediction?

b. For the prediction in part (a), find a 95 percent prediction interval for the unobserved Y value.

21. Use the data in Table 11.6 on page 707. You should perform the least-squares regression requested in Exercise 18 in Sec. 11.2 before starting this exercise.

a. Plot the residuals from the least-squares regression against the 1970 price. Do you see a pattern?

b. Transform both prices to their natural logarithms and repeat the least-squares regression. Now plot the residuals against logarithm of 1970 price. Does this plot look any better than the one in part (a)?

22. Perform a least-squares regression of the logarithm of the 1980 fish price on the 1970 fish price, using the raw data in Table 11.6 on page 707.

a. Test the null hypothesis that the slope β_1 is less than 2.0 at level $\alpha_0 = 0.01$.

b. Find a 90 percent confidence interval for the slope β_1.

c. Find a 90 percent prediction interval for the 1980 price of a species that cost 21.4 in 1970. (Note that 21.4 is the 1970 price, *not* the logarithm of the 1970 price.)

23. Prove that the first test in Theorem 11.3.4 does indeed have level α_0. *Hint:* Use an argument similar to that used to prove part (ii) of Theorem 9.5.1.

24. Find explicit formulas (no sup or inf) for the endpoints of the interval in Eq. (11.3.36) for the following special cases:

a. $c_0 = 1$ and $c_1 = x > 0$.

b. $c_0 = 1$ and $c_1 = x < 0$.

Hint: In both cases the endpoints are of the form $\hat{\beta}_0 + \hat{\beta}_1 x$ plus or minus linear functions of x that depend on the lengths of the sides of the rectangular joint confidence set.

25. In this problem, we will construct a narrower confidence band for a regression function using Theorem 11.3.7. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the least-squares estimators, and let σ' be the estimator of σ used in this section. Let $x_0 < x_1$ be two possible values of the predictor X.

a. Find formulas for the simultaneous coefficient $1 - \alpha_0$ confidence intervals for $\beta_0 + \beta_1 x_0$ and $\beta_0 + \beta_1 x_1$.

b. For each real number x, find the formula for the unique α such that $x = \alpha x_0 + (1 - \alpha)x_1$. Call that value $\alpha(x)$.

c. Call the intervals found in part (a) (A_0, B_0) and (A_1, B_1), respectively. Define the event

$$C = \{A_0 < \beta_0 + \beta_1 x_0 < B_0 \text{ and } A_1 < \beta_0 + \beta_1 x_1 < B_1\}.$$

For each real x, define $L(x)$ and $U(x)$ to be, respectively, the smallest and largest of the following four numbers:

$$\alpha(x)A_0 + [1 - \alpha(x)]A_1,\ \alpha(x)B_0 + [1 - \alpha(x)]A_1,$$
$$\alpha(x)A_0 + [1 - \alpha(x)]B_1,\ \alpha(x)B_0 + [1 - \alpha(x)]B_1.$$

If the event C occurs, prove that, for every real x, $L(x) < \beta_0 + \beta_1 x < U(x)$.

★ 11.4 Bayesian Inference in Simple Linear Regression

In Sec. 8.6, we introduced an improper prior distribution for the mean μ and precision τ of a normal distribution. This prior simplified several calculations associated with the posterior distribution of the parameters. The prior also made some of the resulting inferences bear striking resemblance to inferences based on the sampling distributions of statistics. Something very similar occurs in the simple linear regression setting.

Improper Priors for Regression Parameters

Example 11.4.1 Gasoline Mileage. Once again, consider Example 11.3.2 on page 714. Suppose that we are interested in saying something about how far we think β_1 is from 0 and how strongly we believe that. For example, suppose that we would like to be able to say how likely it is that $|\beta_1|$ is at most c for arbitrary values of c. To do this requires us to compute a distribution for β_1. The posterior distribution of β_1 given the observed data would serve this purpose. ◀

We shall continue to assume that we will observe pairs of variables (X_i, Y_i) for $i = 1, \dots, n$. We shall also assume that the conditional distribution of $Y_1, \dots, Y_n$, given $X_1 = x_1, \dots, X_n = x_n$ and parameters β_0, β_1, and σ^2, is that the Y_i are independent with Y_i having the normal distribution with mean $\beta_0 + \beta_1 x_i$ and variance σ^2. Let $\tau = 1/\sigma^2$ be the precision, as we did in Sec. 8.6. If we let the parameters have an improper prior with "p.d.f." $\xi(\beta_0, \beta_1, \tau) = 1/\tau$, then it is not difficult to find the posterior distribution of the parameters.

Theorem 11.4.1 Suppose that $Y_1, \dots, Y_n$ are independent given $x_1, \dots, x_n$ and β_0, β_1, and τ, with Y_i having the normal distribution with mean $\beta_0 + \beta_1 x_i$ and precision τ. Let the prior distribution be improper with "p.d.f." $\xi(\beta_0, \beta_1, \tau) = 1/\tau$. Then the posterior distribution of β_0, β_1, and τ is as follows. Conditional on τ, the joint distribution of β_0 and β_1 is the bivariate normal distribution with correlation $-n\bar{x}/(n\sum_{i=1}^n x_i^2)^{1/2}$

Table 11.10 Posterior means and variances for simple linear regression with improper prior

Parameter	Mean	Variance
β_0	$\hat{\beta}_0$	$(\frac{1}{n} + \overline{x}^2/s_x^2)/\tau$
β_1	$\hat{\beta}_1$	$(s_x^2\tau)^{-1}$

Table 11.11 Relation between Eq. (5.10.2) and Theorem 11.4.1

(5.10.2)	Theorem 11.4.1
ρ	$-n\overline{x}/(n\sum_{i=1}^n x_i^2)^{1/2}$
σ_1^2	$(\frac{1}{n} + \overline{x}^2/s_x^2)/\tau$
σ_2^2	$(s_x^2\tau)^{-1}$
x_1	β_0
μ_1	$\hat{\beta}_0$
x_2	β_1
μ_2	$\hat{\beta}_1$

and means and variances as given in Table 11.10. The posterior distribution of τ is the gamma distribution with parameters $(n-2)/2$ and $S^2/2$, where S^2 is defined in Eq. (11.3.9). The marginal posterior distribution of

$$\left[\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2}\right]^{-1/2} \frac{c_0\beta_0 + c_1\beta_1 - [c_0\hat{\beta}_0 + c_1\hat{\beta}_1]}{\sigma'} \tag{11.4.1}$$

is the t distribution with $n-2$ degrees of freedom if c_0 and c_1 are not both 0.

Proof The posterior p.d.f. is proportional to the product of the prior p.d.f. and the likelihood function. The likelihood is the conditional p.d.f. of the data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ given the parameters (and $\boldsymbol{x} = (x_1, \ldots, x_n)$), namely,

$$f_n(\boldsymbol{y}|\beta_0, \beta_1, \tau, \boldsymbol{x}) = (\tau/[2\pi])^{n/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right). \tag{11.4.2}$$

To show that the posterior distribution is as stated in the theorem, it suffices to prove that $1/\tau$ times (11.4.2) is proportional (as a function of β_0, β_1, and τ) to the proposed posterior p.d.f.

The proposed posterior p.d.f. of τ is proportional (as a function of τ) to

$$\tau^{(n-2)/2-1}e^{-S^2\tau/2}. \tag{11.4.3}$$

The proposed conditional posterior p.d.f. of (β_0, β_1) given τ is the bivariate normal p.d.f. in Eq. (5.10.2) on page 338 with the substitutions in Table 11.11.

The key to simplifying the substitutions in Eq. (5.10.2) is to note that

$$1-\rho^2=\frac{s_x^2}{\sum_{i=1}^n x_i^2},\ \sigma_1^2=\frac{\sum_{i=1}^n x_i^2}{ns_x^2\tau},\ \text{and}\ \frac{\rho}{\sigma_1\sigma_2}=-\frac{n\overline{x}s_x^2}{\tau\sum_{i=1}^n x_i^2}.$$

The substitutions in Table 11.11 show that the proposed conditional posterior for (β_0, β_1) given τ is proportional to

$$\tau\exp\left(-\frac{\tau}{2}\left[n(\beta_0-\hat{\beta}_0)^2+2n\overline{x}(\beta_0-\hat{\beta}_0)(\beta_1-\hat{\beta}_1)+\left(\sum_{i=1}^n x_i^2\right)(\beta_1-\hat{\beta}_1)^2\right]\right). \tag{11.4.4}$$

The product of (11.4.3) and (11.4.4) is the proposed joint posterior p.d.f., and it is proportional to

$$\tau^{n/2-1}\exp\left(-\frac{\tau}{2}\left[S^2+n(\beta_0-\hat{\beta}_0)^2+2n\overline{x}(\beta_0-\hat{\beta}_0)(\beta_1-\hat{\beta}_1)+\left(\sum_{i=1}^n x_i^2\right)(\beta_1-\hat{\beta}_1)^2\right]\right). \tag{11.4.5}$$

We shall now show that $1/\tau$ times the right side of Eq. (11.4.2) is proportional to (11.4.5). The summation in the exponent of Eq. (11.4.2) is exactly the same as the summation in Eq. (11.3.29) if we remove the asterisks from (11.3.29). In Sec. 11.3, we rewrote (11.3.29) as

$$S^2+n(\beta_0-\hat{\beta}_0)^2+\left(\sum_{i=1}^n x_i^2\right)(\beta_1-\hat{\beta}_1)^2+2n\overline{x}(\beta_0-\hat{\beta}_0)(\beta_1-\hat{\beta}_1), \tag{11.4.6}$$

where the asterisks have been removed from (11.4.6). Notice that (11.4.6) is the same as the factor in the exponent of (11.4.5) that is multiplied by $-\tau^2/2$. Also, notice that $1/\tau$ times the factor multiplying the exponential in (11.4.2) equals $\tau^{n/2-1}$. It follows that $1/\tau$ times (11.4.2) is proportional to (11.4.5).

Finally, we prove that the random variable in (11.4.1) has the t distribution with $n-2$ degrees of freedom. Since (β_0, β_1) has a bivariate normal distribution conditional on τ, it follows that $c_0\beta_0+c_1\beta_1$ has a normal distribution conditional on τ. Its mean is $c_0\hat{\beta}_0+c_1\hat{\beta}_1$. Its variance (given τ) is obtained from Eq. (5.10.9) and Table 11.10 (after some tedious algebra) as v/τ where

$$v=\frac{c_0^2}{n}+c_0^2\frac{\overline{x}^2}{s_x^2}+c_1^2\frac{1}{s_x^2}-2c_0c_1\frac{\overline{x}}{s_x^2}=\frac{c_0^2}{n}+\frac{(c_0\overline{x}-c_1)^2}{s_x^2}.$$

Define the random variable

$$Z=\left(\frac{\tau}{v}\right)^{1/2}(c_0\beta_0+c_1\beta_1-[c_0\hat{\beta}_0+c_1\hat{\beta}_1]),$$

and notice that Z has the standard normal distribution given τ and hence is independent of τ. The distribution of $W=S^2\tau$ is the gamma distribution with parameters $(n-2)/2$ and $1/2$, which is also the χ^2 distribution with $n-2$ degrees of freedom. It follows from the definition of the t distribution that $Z/(W/[n-2])^{1/2}$ has the t distribution with $n-2$ degrees of freedom. Since $\sigma'^2=S^2/(n-2)$, it is straightforward to verify that $Z/(W/[n-2])^{1/2}$ is the same as the random variable in (11.4.1). ■

Example 11.4.2 **Pressure and the Boiling Point of Water.** At the end of Example 11.3.6, we estimated the coefficients of the regression of log-pressure on the boiling point using only 16 of the 17 observations in Forbes' original data. We obtained $\hat{\beta}_0 = -0.9518$ and $\hat{\beta}_1 = 0.0205$ with $\sigma' = 2.616 \times 10^{-3}$. With one observation removed, we have $n = 16$, $\overline{x} = 202.85$, and $s_x^2 = 527.9$. We can now apply Theorem 11.4.1 to make an inference based on the posterior distributions of the parameters. For example, suppose that we are interested in an interval estimate of β_1. Letting $c_0 = 0$ and $c_1 = 1$ in (11.4.1), we find that the posterior distribution of

$$\frac{s_x}{\sigma'}(\beta_1 - \hat{\beta}_1) = 449.2(\beta_1 - 0.0205) \tag{11.4.7}$$

is the t distribution with 15 degrees of freedom. If we want our interval to contain a portion of the posterior distribution with probability $1 - \alpha_0$, then we can note that the posterior probability is $1 - \alpha_0$ that $|449.2(\beta_1 - 0.0205)| \le T_{14}^{-1}(1 - \alpha_0/2)$. For example, if $\alpha_0 = 0.1$, then $T_{14}^{-1}(1 - 0.1/2) = 1.761$. The interval estimate is then $0.0205 \pm 1.761/449.2 = (0.0166, 0.0244)$. ◀

The reader should note that the random variable in Eq. (11.4.7) is the same as U_1 in Eq. (11.3.22) when $\beta_1 = \beta_1^*$. This implies that a coefficient $1 - \alpha_0$ confidence interval for β_1 will be the same as an interval containing posterior probability $1 - \alpha_0$ when we use the improper prior in Theorem 11.4.1. Indeed, the random variable in (11.4.1) is the same as U_{01} in Eq. (11.3.14) for all c_0 and c_1 so long as $c_0\beta_0 + c_1\beta_1 = c_*$. This implies that coefficient $1 - \alpha_0$ confidence intervals for all linear combinations of the regression parameters will also contain probability $1 - \alpha_0$ of the posterior distribution when the improper prior in Theorem 11.4.1 is used. The reader can prove these claims in Exercises 1 and 2 in this section.

Note: There is a Conjugate Family of Proper Prior Distributions. The posterior distribution of the parameters given in Theorem 11.4.1 has the following form: τ has a gamma distribution, and, conditional on τ, (β_0, β_1) has a bivariate normal distribution with variances and covariances that are multiples of $1/\tau$. The collection of distributions of the form just described is a conjugate family of prior distributions for the parameters of simple linear regression. Readers interested in the details of using such priors can consult a text like Broemeling (1985).

Prediction Intervals

On page 716, we showed how to form intervals for predicting future observations. In the Bayesian framework, we can also form intervals for predicting future observations. Let Y be a future observation with corresponding predictor x. Then $Z_1 = \tau^{1/2}(Y - \beta_0 - \beta_1 x)$ has the standard normal distribution conditional on the parameters and the data; hence, it is independent of the parameters and the data. Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ as we did on page 716. It can be shown that the conditional distribution of $Z_2 = \tau^{1/2}(\beta_0 + \beta_1 x - \hat{Y})$ given τ, and the data is the normal distribution with mean 0 and variance

$$\frac{1}{n} + \frac{(x - \overline{x})^2}{s_x^2},$$

and hence it is independent of τ and the data. (See Exercise 3.) Since Z_1 is independent of all of the parameters, it is independent of Z_2, also. It follows that the conditional distribution of $Z_1 + Z_2 = \tau^{1/2}(Y - \hat{Y})$, given τ and the data, is the normal

distribution with mean 0 and variance

$$1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{s_x^2}.$$

As in the proof of Theorem 11.4.1, $S^2\tau$ has the χ^2 distribution with $n-2$ degrees of freedom and is independent of $Z_1 + Z_2$. It follows from the definition of the t distribution that the random variable

$$U_x = \frac{Y - \hat{Y}}{\sigma' \left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{s_x^2}\right]^{1/2}}$$

has the t distribution with $n-2$ degrees of freedom given the data. Hence, the conditional probability, given the data, is $1 - \alpha_0$ that Y is in the interval with endpoints

$$\hat{Y} \pm T_{n-2}^{-1}\left(1 - \frac{\alpha_0}{2}\right)\sigma'\left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{s_x^2}\right]^{1/2}. \qquad (11.4.8)$$

Notice that the U_x defined above is identical to the U_x defined in Eq. (11.3.26). Also, the interval (11.4.8) is the same as the one given in Eq. (11.3.25). The interpretation of the prediction interval based on the posterior distribution is somewhat simpler than the interpretation given after (11.3.25) because the probability is conditional on all of the known quantities (that is, the data). The probability only concerns the distribution of the unknown quantity Y conditional on the data.

Example 11.4.3

Pressure and the Boiling Point of Water. Suppose that we are interested in predicting pressure when the boiling point of water is 208 degrees. We shall find an interval such that the posterior probability is 0.9 that the pressure will be in the interval. That is, we shall use Eq. (11.4.8) with $\alpha_0 = 0.1$ and $x = 208$. We can find $T_{14}(0.95) = 1.761$ from the table of the t distribution in this book. The rest of the necessary values are given in Example 11.4.2. In particular, with Y standing for log-pressure, $\hat{Y} = -0.9518 + 0.0205 \times 208 = 3.3122$, and

$$\sigma'\left[1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{s_x^2}\right]^{1/2} = 2.616 \times 10^{-3}\left[1 + \frac{1}{16} + \frac{(208 - 202.85)^2}{527.9}\right]^{1/2}$$
$$= 2.759 \times 10^{-3}.$$

So our interval for log-pressure has endpoints $3.3122 \pm 1.761 \times 2.759 \times 10^{-3}$, which are 3.307 and 3.317. The interval for pressure itself is then

$$(e^{3.307}, e^{3.317}) = (27.31, 27.58).$$

The reason that we can convert the interval for log-pressure into the interval for pressure so simply is that $3.307 < Y < 3.317$ if and only if $27.31 < e^Y < 27.58$. So, the posterior probability of the first set of inequalities is the same as the posterior probability of the second set of inequalities. ◀

Tests of Hypotheses

On page 607, we began a discussion of tests based on the posterior distribution. If the cost of type I error is w_0 and the cost of type II error is w_1, we found that the Bayes test was to reject the null hypothesis if the posterior probability of the null hypothesis is less than $w_1/(w_0 + w_1)$. Suppose that we use the improper prior and

that the null hypothesis is $H_0: c_0\beta_0 + c_1\beta_1 = c_*$. Since the posterior distribution of $c_0\beta_0 + c_1\beta_1$ is a continuous distribution, it is clear that the posterior probability of the null hypothesis is 0. For this reason, we shall begin by considering Bayes tests only for one-sided hypotheses. Suppose that the hypotheses of interest are

$$\begin{aligned} H_0 &: \quad c_0\beta_0 + c_1\beta_1 \le c_*, \\ H_1 &: \quad c_0\beta_0 + c_1\beta_1 > c_*. \end{aligned} \tag{11.4.9}$$

The other direction can be handled in a similar fashion. Let $\alpha_0 = w_1/(w_0 + w_1)$. The posterior probability that the null hypothesis is true is the posterior probability that $c_0\beta_0 + c_1\beta_1 \le c_*$. We have already derived the posterior distribution of $c_0\beta_0 + c_1\beta_1$ in Theorem 11.4.1. So, we can compute

$$\begin{aligned} &\Pr(c_0\beta_0 + c_1\beta_1 \le c_*) \\ &= \Pr\left(\left[\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2}\right]^{-1/2} \frac{c_0\beta_0 + c_1\beta_1 - [c_0\hat{\beta}_0 + c_1\hat{\beta}_1]}{\sigma'} \right. \\ &\qquad \left. \le \left[\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2}\right]^{-1/2} \frac{c_* - [c_0\hat{\beta}_0 + c_1\hat{\beta}_1]}{\sigma'}\right) \\ &= T_{n-2}\left(\left[\frac{c_0^2}{n} + \frac{(c_0\overline{x} - c_1)^2}{s_x^2}\right]^{-1/2} \frac{c_* - [c_0\hat{\beta}_0 + c_1\hat{\beta}_1]}{\sigma'}\right) \\ &= T_{n-2}(-U_{01}), \end{aligned}$$

where T_{n-2} denotes the c.d.f. of the t distribution with $n-2$ degrees of freedom and U_{01} is the random variable defined in Eq. (11.3.14). It is simple to see that $T_{n-2}(-U_{01}) \le \alpha_0$ if and only if $U_{01} \ge T_{n-2}^{-1}(1-\alpha_0)$. Hence, the Bayes test of the hypotheses (11.4.9) is the same as the level α_0 test of these same hypotheses that was derived after Eq. (11.3.16). Hence, all of the one-sided tests that we learned how to perform in Sec. 11.3 are also Bayes tests when the improper prior is used.

On page 610, we began a discussion of how to deal with two-sided alternatives when the posterior distribution of the parameter was continuous. The same approach can be used in linear regression problems. We shall illustrate with an example.

Example 11.4.4 Gasoline Mileage. In Example 11.4.1, we wanted to make use of the posterior distribution of the slope parameter β_1 from Example 11.3.2 in order to be able to say how likely we believe it is that β_1 is close to 0. We can draw a plot of the posterior c.d.f. of $|\beta_1|$ by making use of Theorem 11.4.1. The posterior distribution of $s_x(\beta_1 - \hat{\beta}_1)/\sigma'$ is the t distribution with $n-2$ degrees of freedom. In Example 11.3.2, we computed $s_x = 1036.9$, $\sigma' = 7.191 \times 10^{-3}$, $\hat{\beta}_1 = 1.396 \times 10^{-4}$, and $n = 173$. It follows that, for all positive c,

$$\begin{aligned} \Pr(|\beta_1| \le c) = \Pr(-c \le \beta_1 \le c) &= T_{171}\left(\frac{1036.9}{7.181 \times 10^{-3}}(c - 1.396 \times 10^{-4})\right) \\ &\quad - T_{171}\left(\frac{1036.9}{7.181 \times 10^{-3}}(-c - 1.396 \times 10^{-4})\right), \end{aligned}$$

where T_{171} is the c.d.f. of the t distribution with 171 degrees of freedom. Figure 11.14 contains a plot of the posterior c.d.f. of $|\beta_1|$. We can see that the probability is essentially 1 that $|\beta_1| < 1.6 \times 10^{-4}$, but it is also essentially 1 that $|\beta_1| > 1.2 \times 10^{-4}$. These numbers may look small. However, remember that β_1 must get multiplied by

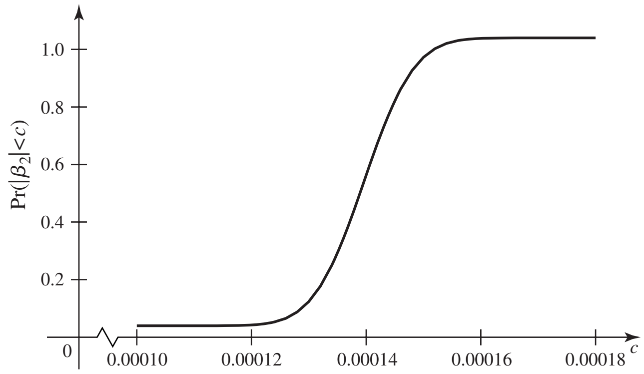


Figure 11.14 Plot of posterior c.d.f. of $|\beta_1|$ in Example 11.4.4.

horsepower, which is typically a number in the 50–300 range. So, even if β_1 is as small as 1.2×10^{-4}, the difference between gallons per mile at 100 and 200 horsepower will be 0.012, which is a sizeable difference in gallons per mile. We can also translate this result into miles per gallon. Suppose that $\beta_1 = 1.2 \times 10^{-4}$, and suppose that β_0 equals its conditional mean given that $\beta_1 = 1.2 \times 10^{-4}$. This conditional mean can be computed using the method of Exercise 7, and it equals 0.01897. Then the miles per gallon for a 200 horsepower car is 23.27, and the miles per gallon for a 100 horsepower car is 32.23. ◀

Summary

We have used improper prior distributions for the parameters of the simple linear regression model, and we have found the posterior distributions of the parameters after observing n data points. The posterior distributions of the intercept and slope parameters are t distributions with $n - 2$ degrees of freedom that have been shifted and rescaled. These posterior distributions show striking similarities to the sampling distributions of the least-squares estimators. Indeed, posterior probability intervals for the parameters are exactly the same as confidence intervals, prediction intervals for future observations are the same as those based on the sampling distributions, and level α_0 tests of one-sided null and alternative hypotheses reject the null hypotheses when the posterior probability of the null hypothesis is less than α_0. The only significant lack of connection between posterior calculations and those based on sampling distributions is the testing of hypotheses in which the alternative is two-sided.

Exercises

1. Assume the usual conditions for simple linear regression. Assume that we use the improper prior discussed in this section. Let (a, b) be the observed value of a coefficient $1 - \alpha_0$ confidence interval for β_1 constructed as in Sec. 11.3. Prove that the posterior probability is $1 - \alpha_0$ that $a < \beta_1 < b$.

2. Assume the usual conditions for simple linear regression. Assume that we use the improper prior discussed in this section. Let (a, b) be the observed value of a coefficient $1 - \alpha_0$ confidence interval for $c_0\beta_0 + c_1\beta_1$ constructed as in Sec. 11.3. Prove that the posterior probability is $1 - \alpha_0$ that $a < c_0\beta_0 + c_1\beta_1 < b$.

3. Assume a simple linear regression model with the improper prior. Show that, conditional on τ, the posterior distribution of $\tau^{1/2}(\beta_0 + \beta_1 x - \hat{Y})$ is the normal distribution with mean 0 and variance

$$\frac{1}{n} + \frac{(x - \overline{x})^2}{s_x^2}.$$

4. We wish to fit a simple linear regression model to the data in Table 11.9 on page 727. Use an improper prior distribution.

a. Find the posterior distribution of the parameters.

b. Find a bounded interval that contains 90 percent of the posterior distribution of β_1.

c. Find the probability that β_0 is between 0 and 2.

5. Use the data in Table 11.9, and suppose that we wish to fit a simple linear regression model to the data. Use the improper prior.

a. Find the posterior distribution of the slope parameter β_1.

b. Find the posterior distribution of $\beta_0 + \beta_1$, the mean of a future observation Y corresponding to $x = 1$.

c. Draw a graph of the posterior c.d.f. of $|\beta_1 - 0.7|$.

6. Use the data in Table 11.6 on page 707. Assume that we wish to fit a simple linear regression model for predicting logarithm of 1980 price from logarithm of 1970 price.

a. Find the posterior distribution of the slope parameter β_1.

b. Find the posterior probability that $\beta_1 \le 2$.

c. Find a 95 percent prediction interval for the 1980 price of a species that cost 21.4 in 1970.

7. In a simple linear regression problem with the usual improper prior, prove that the conditional mean of β_0 given β_1 is $\hat{\beta}_0 - \overline{x}(\beta_1 - \hat{\beta}_1)$. *Hint:* Use the fact that (β_0, β_1) has a bivariate normal distribution as described in Theorem 11.4.1, and then use Eq. (5.10.6) to find the conditional mean.

11.5 The General Linear Model and Multiple Regression

The simple linear regression model can be extended to allow the mean of Y to be a function of several predictor variables. Much of the resulting distribution theory, is very similar to the simple regression case.

The General Linear Model

Example 11.5.1 Unemployment in the 1950s. The data in Table 11.12 provide the unemployment rates during the 10 years from 1950 to 1959 together with an index of industrial production from the Federal Reserve Board. It might make sense to think that unemployment is related to industrial production. Other factors also play a role, and those other factors most likely changed over the course of the decade. As a surrogate for these other factors, some function of the year could be included as a predictor. Figure 11.15 shows plots of unemployment against each of the two predictor variables. It is not clear from the plots precisely how unemployment varies with the two predictors, but there appear to be some relationships. In this section, we shall show how to fit a regression model with more than one predictor to these and other data. ◀

In this section, we shall study regression problems in which the observations $Y_1, \dots, Y_n$ satisfy assumptions like Assumptions 11.2.1–11.2.5 that were made in Sections 11.2 and 11.3. In particular, we shall again assume that each observation Y_i has a normal distribution, that the observations $Y_1, \dots, Y_n$ are independent, and that the observations $Y_1, \dots, Y_n$ have the same variance σ^2. Instead of a single predictor being associated with each Y_i, we assume that a p-dimensional vector $z_i = (z_{i0}, \dots, z_{ip-1})$ is associated with each Y_i. The assumptions that we make can now be restated in this framework.

Assumption 11.5.1 Predictor is known. Either the vectors $z_1, \dots, z_n$ are known ahead of time, or they are the observed values of random vectors $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$ on whose values we condition before computing the joint distribution of $(Y_1, \dots, Y_n)$.

Table 11.12 Unemployment data for Example 11.5.1

Unemployment	Index of production	Year
3.1	113	1950
1.9	123	1951
1.7	127	1952
1.6	138	1953
3.2	130	1954
2.7	146	1955
2.6	151	1956
2.9	152	1957
4.7	141	1958
3.8	159	1959

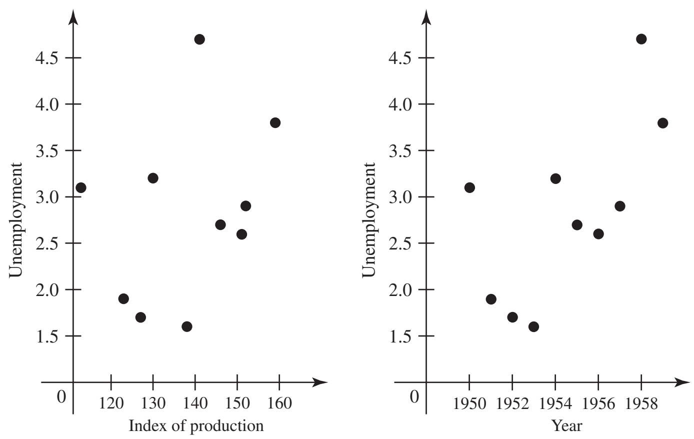


Figure 11.15 Plots of unemployment against the two predictor variables for Example 11.5.1.

Assumption 11.5.2 Normality. For $i = 1, \ldots, n$, the conditional distribution of Y_i given the vectors $z_1, \ldots, z_n$ is a normal distribution.

Assumption 11.5.3 Linear Mean. There is a vector of parameters $\beta = (\beta_0, \ldots, \beta_{p-1})$ such that the conditional mean of Y_i given the values $z_1, \ldots, z_n$ has the form

$$z_{i0}\beta_0 + z_{i1}\beta_1 + \cdots + z_{ip-1}\beta_{p-1}, \tag{11.5.1}$$

for $i = 1, \ldots, n$.

Assumption 11.5.4 Common Variance. There is a parameter σ^2 such that the conditional variance of Y_i given the values $z_1, \ldots, z_n$ is σ^2 for $i = 1, \ldots, n$.

Assumption 11.5.5 Independence. The random variables $Y_1, \ldots, Y_n$ are independent given the observed $z_1, \ldots, z_n$.

The generalization that we introduce here is that the mean of each observation Y_i is a linear combination of p unknown parameters $\beta_0, \ldots, \beta_{p-1}$ as in (11.5.1). Each value z_{ij} either may be fixed by the experimenter before the experiment is started or may be observed in the experiment along with the value of Y_i. In the latter case, Eq. (11.5.1) gives the conditional mean of Y_i given the observed z_{ij} values.

Definition 11.5.1 General Linear Model. The statistical model in which the observations $Y_1, \ldots, Y_n$ satisfy Assumptions 11.5.1–11.5.5 is called the *general linear model*.

In Definition 11.5.1, the term *linear* refers to the fact that the expectation of each observation Y_i is a linear function of the unknown parameters $\beta_0, \ldots, \beta_{p-1}$.

Many different types of regression problems are examples of general linear models. For example, in a problem of simple linear regression, $E(Y_i) = \beta_0 + \beta_1 x_i$ for $i = 1, \ldots, n$. This expectation can be represented in the form given in Eq. (11.5.1), with $p = 2$, by letting $z_{i0} = 1$ and $z_{i1} = x_i$ for $i = 1, \ldots, n$. Similarly, if the regression of Y on X is a polynomial of degree k, then, for $i = 1, \ldots, n$,

$$E(Y_i) = \beta_0 + \beta_1 x_i + \cdots + \beta_k x_i^k. \tag{11.5.2}$$

In this case, $p = k + 1$ and $E(Y_i)$ can be represented in the form given in Eq. (11.5.1) by letting $z_{ij} = x_i^j$ for $j = 0, \ldots, k$.

As a final example, consider a problem in which the regression of Y on k variables $X_1, \ldots, X_k$ is a function like that given in Eq. (11.2.1). A problem of this type is called a problem of *multiple linear regression* because we are considering the regression of Y on k variables $X_1, \ldots, X_k$, rather than on just a single variable X, and we are assuming also that this regression is a linear function of the parameters $\beta_0, \ldots, \beta_k$. In a problem of multiple linear regression, we obtain n vectors of observations $(x_{i1}, \ldots, x_{ik}, Y_i)$, for $i = 1, \ldots, n$. Here x_{ij} is the observed value of the variable X_j for the ith observation. Then $E(Y_i)$ is given by the relation

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}. \tag{11.5.3}$$

This expectation can also be represented in the form given in Eq. (11.5.1), with $p = k + 1$, by letting $z_{i0} = 1$ and $z_{ij} = x_{ij}$ for $j = 1, \ldots, k$.

Example 11.5.2 Unemployment in the 1950s. In Example 11.5.1, we can let Y stand for the unemployment rate, while X_1 stands for the index of production and X_2 stands for the year. ◄

Our discussion has indicated that the general linear model is general enough to include problems of simple and multiple linear regression, problems in which the regression function is a polynomial, problems in which the regression function has the form given in Eq. (11.1.16), and many other problems.

Some books devoted to regression and other linear models are Cook and Weisberg (1999), Draper and Smith (1998), Graybill and Iyer (1994), and Weisberg (1985).

Maximum Likelihood Estimators

We shall now describe a procedure for determining the M.L.E.'s of $\beta_0, \ldots, \beta_{p-1}$ in the general linear model. Since $E(Y_i)$ is given by Eq. (11.5.1) for $i = 1, \ldots, n$, the likelihood function after observing values $y_1, \ldots, y_n$ will have the following form:

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - z_{i0}\beta_0 - \cdots - z_{ip-1}\beta_{p-1})^2\right]. \tag{11.5.4}$$

Since the M.L.E.'s are the values that maximize the likelihood function (11.5.4), it can be seen that the estimates $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$ will be the values of $\beta_0, \dots, \beta_{p-1}$ for which the following sum of squares Q is minimized:

$$Q = \sum_{i=1}^{n}(y_i - z_{i0}\beta_0 - \cdots - z_{ip-1}\beta_{p-1})^2. \tag{11.5.5}$$

Since Q is the sum of the squares of the deviations of the observed values from the linear function given in Eq. (11.5.1), it follows that the M.L.E.'s $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$ will be the same as the least-squares estimates.

To determine the values of $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$, we can calculate the p partial derivatives $\partial Q/\partial \beta_j$ for $j = 0, \dots, p-1$ and can set each of these derivatives equal to 0. The resulting p equations, which are called the *normal equations*, will form a set of p linear equations in $\beta_0, \dots, \beta_{p-1}$. We shall assume that the $p \times p$ matrix formed by the coefficients of $\beta_0, \dots, \beta_{p-1}$ in the normal equations is nonsingular. Then these equations will have a unique solution $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$, and $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$ will be both the M.L.E.'s and the least-squares estimates of $\beta_0, \dots, \beta_{p-1}$.

For a problem of polynomial regression in which $E(Y_i)$ is given by Eq. (11.5.2), the normal equations were presented as the relations (11.1.8). For a problem of multiple linear regression in which $E(Y_i)$ is given by Eq. (11.5.3), the normal equations were presented as the relations (11.1.13).

If we substitute $\hat{\beta}_i$ for β_i for $i = 0, \dots, p-1$ in the formula for Q in Eq. (11.5.5), we obtain

$$S^2 = \sum_{i=1}^{n}(Y_i - z_{i0}\hat{\beta}_0 - \cdots - z_{ip-1}\hat{\beta}_{p-1})^2. \tag{11.5.6}$$

Eq. (11.5.6) is the natural generalization of Eq. (11.3.9) to the multiple regression case. It can be shown using the same method outlined in the proof of Theorem 11.2.1 that the M.L.E. of σ^2 in the general linear model is

$$\hat{\sigma}^2 = \frac{S^2}{n}. \tag{11.5.7}$$

The details are left to Exercise 1 at the end of this section. In analogy to Eq. (11.3.12), we define the useful quantity

$$\sigma' = \left(\frac{S^2}{n-p}\right)^{1/2}. \tag{11.5.8}$$

This makes σ'^2 an unbiased estimator of σ^2. (See Exercise 2.)

Explicit Form of the Estimators

In order to derive the explicit form and the properties of the estimators $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$, it is convenient to use the notation and techniques of vectors and matrices. We shall let the $n \times p$ matrix $\boldsymbol{Z}$ be defined as follows:

$$\boldsymbol{Z} = \begin{bmatrix} z_{10} & \cdots & z_{1p-1} \\ z_{20} & \cdots & z_{2p-1} \\ \vdots & \ddots & \vdots \\ z_{n0} & \cdots & z_{np-1} \end{bmatrix}. \tag{11.5.9}$$

This matrix $\mathbf{Z}$ distinguishes one regression problem from another, because the entries in $\mathbf{Z}$ determine the particular linear combinations of the unknown parameters $\beta_0, \ldots, \beta_{p-1}$ that are relevant in a given problem.

Definition 11.5.2 Design Matrix. The matrix $\mathbf{Z}$ in Eq. (11.5.9) for a general linear model is called the *design matrix* of the model.

The name "design matrix" comes from the case in which the z_{ij} are chosen by the experimenter to achieve a well-designed experiment. It should be kept in mind, however, that some or all of the entries in $\mathbf{Z}$ may be simply the observed values of certain variables, and may not actually be controlled by the experimenter.

We shall also let $\boldsymbol{y}$ be the $n \times 1$ vector of observed values of $Y_1, \ldots, Y_n$, $\boldsymbol{\beta}$ be the $p \times 1$ vector of parameters, and $\hat{\boldsymbol{\beta}}$ be the $p \times 1$ vector of estimates. These vectors may be represented as follows:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \quad \text{and} \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}.$$

The transpose of a vector or matrix $\boldsymbol{v}$ will be denoted by $\boldsymbol{v}'$.

Theorem 11.5.1 General Linear Model Estimators. The least squares estimator (and M.L.E.) of β is

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}. \tag{11.5.10}$$

Proof The sum of squares Q given in Eq. (11.5.5) can be written in the following form:

$$Q = (\boldsymbol{y} - \boldsymbol{Z\beta})'(\boldsymbol{y} - \boldsymbol{Z\beta}).$$

Since Q is a quadratic function of the coordinates of $\boldsymbol{\beta}$, it is straightforward to take the partial derivatives of Q with respect to these coordinates and set them equal to 0. For example, the partial derivative with respect to β_0 is

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n} z_{i0}y_i + 2\sum_{j=0}^{p-1}\beta_j \sum_{i=1}^{n} z_{i0}z_{ij}. \tag{11.5.11}$$

Each of the other partial derivatives produces an equation similar to (11.5.11). Set the right-hand sides of each of these p equations to 0, and arranged them into the following matrix equation:

$$\boldsymbol{Z}'\boldsymbol{Z\beta} = \boldsymbol{Z}'\boldsymbol{y}. \tag{11.5.12}$$

Because it is assumed that the $p \times p$ matrix $\mathbf{Z}'\mathbf{Z}$ is nonsingular, the vector of estimates $\hat{\boldsymbol{\beta}}$ will be the unique solution of Eq. (11.5.12). In order for $\mathbf{Z}'\mathbf{Z}$ to be nonsingular, the number of observations n must be at least p, and there must be at least p linearly independent rows in the matrix $\mathbf{Z}$. When this assumption is satisfied, it follows from Eq. (11.5.12) that $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$. Thus, if we replace the vector $\boldsymbol{y}$ of observed values by the vector $\boldsymbol{Y}$ of random variables, the form for the vector of estimators $\hat{\boldsymbol{\beta}}$ will be (11.5.10). ■

Virtually every statistical computer package will calculate least-squares estimates for a multiple linear regression. Even some handheld calculators will perform multiple linear regression. The matrix $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ is useful for more than just computing $\hat{\boldsymbol{\beta}}$ in

Eq. (11.5.10), as we shall see later in this section. Not every piece of regression software makes it easy to access this matrix.

It follows from Eq. (11.5.10) that each of the estimators $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$ will be a linear combination of the coordinates $Y_1, \dots, Y_n$ of the vector $\boldsymbol{Y}$. Since each of these coordinates has a normal distribution and they are independent, it follows that each estimator $\hat{\beta}_j$ will also have a normal distribution. Indeed, the entire vector $\hat{\boldsymbol{\beta}}$ has a joint normal distribution (called a *multivariate normal distribution*), which is a generalization of the bivariate normal distribution to more than two coordinates. We shall not discuss the multivariate normal distribution in detail in this text, but we shall merely point out one feature that it has in common with the bivariate normal distribution: If a vector $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution, then every linear combination of the coordinates of $\hat{\boldsymbol{\beta}}$ has a normal distribution. Indeed, every collection of linear combinations of the coordinates of $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution.

Example 11.5.3 Unemployment in the 1950s. The matrix $\boldsymbol{Z}$ in Example 11.5.1 has three columns. The first column is the number 1 ten times. The second column is the second column of Table 11.12. In order to avoid some numerical problems, we shall let the third column of $\boldsymbol{Z}$ be the third column of Table 11.12 minus 1949. The vector $\boldsymbol{y}$ is the first column of Table 11.12. We can then compute the matrix $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ and the vector $\boldsymbol{Z}'\boldsymbol{y}$:

$$(\boldsymbol{Z}'\boldsymbol{Z})^{-1} = \begin{pmatrix} 38.35 & -0.3323 & 1.383 \\ -0.3323 & 2.915 \times 10^{-3} & -0.01272 \\ 1.383 & -0.01272 & 0.06762 \end{pmatrix} \qquad \boldsymbol{Z}'\boldsymbol{y} = \begin{pmatrix} 28.2 \\ 3931 \\ 144.1 \end{pmatrix}.$$

We can then use Eq. (11.5.10) to compute

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 13.45 \\ -0.1033 \\ 0.6594 \end{pmatrix}.$$

We shall examine the residuals later in this section. ◄

Mean Vector and Covariance Matrix

We shall now derive the means, variances, and covariances of $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$. Suppose that $\boldsymbol{Y}$ is an n-dimensional random vector with coordinates $Y_1, \dots, Y_n$. Thus,

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}. \tag{11.5.13}$$

The expectation $E(\boldsymbol{Y})$ of this random vector is defined to be the n-dimensional vector whose coordinates are the expectations of the individual coordinates of $\boldsymbol{Y}$. Hence,

$$E(\boldsymbol{Y}) = \begin{bmatrix} E(Y_1) \\ \vdots \\ E(Y_n) \end{bmatrix}.$$

Definition 11.5.3 Mean Vector/Covariance Matrix. If $\boldsymbol{Y}$ is a random vector, then the vector $E(\boldsymbol{Y})$ is called the *mean vector* of $\boldsymbol{Y}$. The *covariance matrix* of $\boldsymbol{Y}$ is defined to be the $n \times n$ matrix such that, for $i = 1, \dots, n$ and $j = 1, \dots, n$, the element in the ith row and jth column is $\text{Cov}(Y_i, Y_j)$. We shall let $\text{Cov}(\boldsymbol{Y})$ denote this covariance matrix.

For example, if $\text{Cov}(Y_i, Y_j) = \sigma_{ij}$ for all i and j, then

$$\text{Cov}(\boldsymbol{Y}) = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix}.$$

For $i = 1, \dots, n$, $\text{Var}(Y_i) = \text{Cov}(Y_i, Y_i) = \sigma_{ii}$. Therefore, the n diagonal elements of the matrix $\text{Cov}(\boldsymbol{Y})$ are the variances of $Y_1, \dots, Y_n$. Furthermore, since $\text{Cov}(Y_i, Y_j) = \text{Cov}(Y_j, Y_i)$, then $\sigma_{ij} = \sigma_{ji}$. Therefore, the matrix $\text{Cov}(\boldsymbol{Y})$ must be symmetric.

The mean vector and the covariance matrix of the random vector $\boldsymbol{Y}$ in the general linear model can easily be determined. It follows from Eq. (11.5.1) that

$$E(\boldsymbol{Y}) = \boldsymbol{Z\beta}. \tag{11.5.14}$$

Also, the coordinates $Y_1, \dots, Y_n$ of $\boldsymbol{Y}$ are independent, and the variance of each of these coordinates is σ^2. Therefore,

$$\text{Cov}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}, \tag{11.5.15}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix.

The following result helps us find the mean vector and covariance matrix of $\hat{\boldsymbol{\beta}}$.

Theorem 11.5.2 Suppose that $\boldsymbol{Y}$ is an n-dimensional random vector as specified by Eq. (11.5.13), for which the mean vector $E(\boldsymbol{Y})$ and the covariance matrix $\text{Cov}(\boldsymbol{Y})$ exist. Suppose also that $\boldsymbol{A}$ is a $p \times n$ matrix whose elements are constants, and that $\boldsymbol{W}$ is a p-dimensional random vector defined by the relation $\boldsymbol{W} = \boldsymbol{AY}$. Then $E(\boldsymbol{W}) = \boldsymbol{A}E(\boldsymbol{Y})$ and $\text{Cov}(\boldsymbol{W}) = \boldsymbol{A}\,\text{Cov}(\boldsymbol{Y})\boldsymbol{A}'$.

Proof Let the elements of matrix $\boldsymbol{A}$ be denoted as follows:

$$\boldsymbol{A} = \begin{bmatrix} a_{01} & \cdots & a_{0n} \\ \vdots & \ddots & \vdots \\ a_{p-11} & \cdots & a_{p-1n} \end{bmatrix}.$$

Then the ith coordinate of the vector $E(\boldsymbol{W})$ is

$$E(W_i) = E\left(\sum_{j=1}^{n} a_{ij} Y_j\right) = \sum_{j=1}^{n} a_{ij} E(Y_j). \tag{11.5.16}$$

It can be seen that the final summation in Eq. (11.5.16) is the ith coordinate of the vector $\boldsymbol{A}E(\boldsymbol{Y})$. Hence, $E(\boldsymbol{W}) = \boldsymbol{A}E(\boldsymbol{Y})$.

Next, for $i = 0, \dots, p-1$ and $j = 0, \dots, p-1$, the element in the ith row and jth column of the $p \times p$ matrix $\text{Cov}(\boldsymbol{W})$ is

$$\text{Cov}(W_i, W_j) = \text{Cov}\left(\sum_{r=1}^{n} a_{ir} Y_r, \sum_{s=1}^{n} a_{js} Y_s\right).$$

Therefore, by Exercise 8 of Sec. 4.6,

$$\text{Cov}(W_i, W_j) = \sum_{r=1}^{n} \sum_{s=1}^{n} a_{ir} a_{js}\,\text{Cov}(Y_r, Y_s). \tag{11.5.17}$$

Using the formula for matrix multiplication, one finds that the right side of Eq. (11.5.17) is the element in the ith row and jth column of the $p \times p$ matrix $\boldsymbol{A}\,\text{Cov}(\boldsymbol{Y})\boldsymbol{A}'$. Hence, $\text{Cov}(\boldsymbol{W}) = \boldsymbol{A}\,\text{Cov}(\boldsymbol{Y})\boldsymbol{A}'$. ■

The means, the variances, and the covariances of the estimators $\hat{\beta}_0, \dots, \hat{\beta}_{p-1}$ can be obtained by applying Theorem 11.5.2.

Theorem 11.5.3 In the general linear model, $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

Proof Eq. (11.5.10) says that $\hat{\boldsymbol{\beta}}$ can be represented in the form $\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}$, where $\boldsymbol{A} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$. Therefore, it follows from Theorem 11.5.2 and Eq. (11.5.14) that

$$E(\hat{\boldsymbol{\beta}}) = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'E(\boldsymbol{Y}) = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Also, it follows from Theorem 11.5.2 and Eq. (11.5.15) that

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\beta}}) &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\,\text{Cov}(\boldsymbol{Y})\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\\ &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\sigma^2\boldsymbol{I})\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\\ &= \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}.\end{aligned}$$ ■

Thus, $E(\hat{\beta}_j) = \beta_j$ for $j = 0, \dots, p-1$, and for $j = 1, \dots, n$, $\text{Var}(\hat{\beta}_j)$ equals σ^2 times the jth diagonal entry of the matrix $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$. Also, for $i \neq j$, $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$ will be equal to σ^2 times the entry in the ith row and jth column of the matrix $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

Example 11.5.4 Dishwasher Shipments. The United States Department of Commerce collects data on factory shipments of durable goods as well as other economic indicators. Table 11.13 contains the numbers of factory shipments of dishwashers (in thousands) and private residential investment in billions of 1972 dollars for the years 1960 through 1985. Figure 11.16 shows plots of dishwasher shipments against year and private residential investment. Let Y stand for dishwasher shipments. We could fit a model in which the mean of Y is given by Eq. (11.5.3) with $k = 2$. The matrix $\boldsymbol{Z}$ would have three columns and 26 rows. The first column would be all the number 1. The second column would have time, expressed as the year minus 1960 for numerical stability. The third column would have private residential investment. We can then compute

$$(\boldsymbol{Z}'\boldsymbol{Z})^{-1} = \begin{pmatrix} 1.152 & 0.01279 & -0.02660\\ 0.01279 & 0.001136 & -0.0005636\\ -0.02660 & -0.0005636 & 0.0007026 \end{pmatrix}.$$

The correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ can be computed as

$$\rho = \frac{\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)}{(\text{Var}(\hat{\beta}_1)\,\text{Var}(\hat{\beta}_2))^{1/2}} = \frac{-0.0005636\sigma^2}{(0.001136\sigma^2 \times 0.0007026\sigma^2)^{1/2}} = -0.6309.$$

Notice that the correlation does not depend on the unknown value of σ^2, but only on the design matrix. Also notice that the correlation is negative and sizeable. If one of the coefficients is overestimated, the other one will tend to be underestimated. ◄

Table 11.13 Dishwasher shipments and residential investment from 1960–1985

Year	Dishwasher shipments (thousands)	Private residential investment (billions of 1972 dollars)
1960	555	34.2
1961	620	34.3
1962	720	37.7
1963	880	42.5
1964	1050	43.1
1965	1290	42.7
1966	1528	38.2
1967	1586	37.1
1968	1960	43.1
1969	2118	43.6
1970	2116	41.0
1971	2477	53.7
1972	3199	63.8
1973	3702	62.3
1974	3320	48.2
1975	2702	42.2
1976	3140	51.2
1977	3356	60.7
1978	3558	62.4
1979	3488	59.1
1980	2738	47.1
1981	2484	44.7
1982	2170	37.8
1983	3092	52.7
1984	3491	60.3
1985	3536	61.4

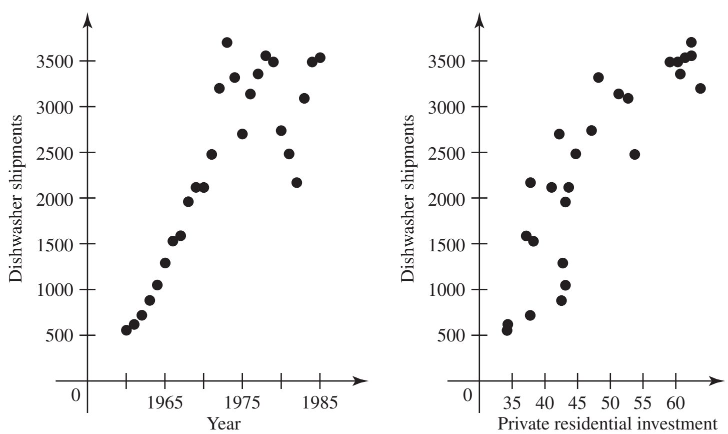


Figure 11.16 Plots of dishwasher shipments against year (left) and private residential investment (right).

The Joint Distribution of the Estimators

Let the random variable S^2 be defined as in Eq. (11.5.6). The sum of squares S^2 can also be represented in the following form:

$$S^2 = (\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}). \tag{11.5.18}$$

The method in the proof of Theorem 11.3.2 can be extended by making use of methods that are beyond the scope of this book in order to prove the following two facts. First, S^2/σ^2 has the χ^2 distribution with $n - p$ degrees of freedom. Second, S^2 and the random vector $\hat{\boldsymbol{\beta}}$ are independent.

From Eq. (11.5.7), we see that $\hat{\sigma}^2 = S^2/n$. Hence, the random variable $n\hat{\sigma}^2/\sigma^2$ has the χ^2 distribution with $n - p$ degrees of freedom, and the estimators $\hat{\sigma}^2$ and $\hat{\boldsymbol{\beta}}$ are independent.

The following result summarizes what we have proven and stated without proof concerning the joint distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$.

Corollary 11.5.1 Let the entries in the symmetric $p \times p$ matrix $(\mathbf{Z}'\mathbf{Z})^{-1}$ be denoted as follows:

$$(\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} \zeta_{00} & \cdots & \zeta_{0p-1} \\ \vdots & \ddots & \vdots \\ \zeta_{p-10} & \cdots & \zeta_{p-1p-1} \end{bmatrix}. \tag{11.5.19}$$

For $j = 0, \ldots, p - 1$, the estimator $\hat{\beta}_j$ has the normal distribution with mean β_j and variance $\zeta_{jj}\sigma^2$. Furthermore, for $i \neq j$, we have $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \zeta_{ij}\sigma^2$. Also, the entire vector $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution. Finally, $\hat{\sigma}^2$ is independent of $\hat{\boldsymbol{\beta}}$ and $n\hat{\sigma}^2/\sigma^2$ has the χ^2 distribution with $n - p$ degrees of freedom. ■

Note that $\hat{\boldsymbol{\beta}}$ is also independent of σ'^2 from Eq. (11.5.8).

Testing Hypotheses

Suppose that it is desired to test the hypothesis that one of the regression coefficients β_j has a particular value β_j^*. In other words, suppose that the following hypotheses are to be tested:

$$\begin{aligned} H_0:&\quad \beta_j = \beta_j^*, \\ H_1:&\quad \beta_j \neq \beta_j^*. \end{aligned} \tag{11.5.20}$$

Since $\text{Var}(\hat{\beta}_j) = \zeta_{jj}\sigma^2$, it follows that when H_0 is true, the following random variable W_j will have the standard normal distribution:

$$W_j = \frac{(\hat{\beta}_j - \beta_j^*)}{\zeta_{jj}^{1/2}\sigma}.$$

Furthermore, since the random variable S^2/σ^2 has the χ^2 distribution with $n - p$ degrees of freedom, and since S^2 and $\hat{\beta}_j$ are independent, it follows that when H_0 is true, the following random variable U_j will have the t distribution with $n - p$ degrees

of freedom:

$$U_j = \frac{W_j}{\left[\frac{1}{n-p}\left(\frac{S^2}{\sigma^2}\right)\right]^{1/2}} = \frac{(\hat{\beta}_j - \beta_j^*)}{(\zeta_{jj})^{1/2}\sigma'}. \tag{11.5.21}$$

The level α_0 test of the hypotheses (11.5.20) specifies that the null hypothesis H_0 should be rejected if $|U_j| \geq T_{n-p}^{-1}(1-\alpha_0/2)$, where T_{n-p}^{-1} is the quantile function of the t distribution with $n-p$ degrees of freedom. Furthermore, if u is the value of U_j observed in a given problem, the corresponding p-value is

$$\Pr(U_j \geq |u|) + \Pr(U_j \leq -|u|). \tag{11.5.22}$$

Tests for one-sided hypotheses can be derived in a similar fashion.

Example 11.5.5

Dishwasher Shipments. In Example 11.5.4, the least-squares estimates for the model are $\hat{\beta}_0 = -1314$, $\hat{\beta}_1 = 66.91$, and $\hat{\beta}_2 = 58.86$. The observed value of σ' is 352.9. Now suppose that we are interested in testing the hypotheses

$$H_0: \quad \beta_1 = 0,$$
$$H_1: \quad \beta_1 \neq 0,$$

where β_1 is the coefficient of time in the multiple linear regression model. Using the matrix $(\mathbf{Z}'\mathbf{Z})^{-1}$ found in Example 11.5.4, we can calculate

$$U_1 = \frac{66.91 - 0}{(0.001136)^{1/2} \times 352.9} = 5.625.$$

The degrees of freedom are $26 - 3 = 23$, and 5.625 is larger than every quantile listed in the table of the t distribution in this book. Using a computer program, we find that the p-value is about 1×10^{-5}. ◀

Example 11.5.6

Unemployment in the 1950s. In Example 11.5.3, we regressed unemployment on a Federal Reserve Board index of production and time. The least-squares estimates are $\hat{\beta}_0 = 13.45$, $\hat{\beta}_1 = -0.1033$, and $\hat{\beta}_2 = 0.6594$. The observed value of σ' is 0.4011. Now suppose that we wish to test the hypotheses

$$H_0: \quad \beta_2 \leq 0.4,$$
$$H_1: \quad \beta_2 > 0.4.$$

To test these hypotheses, we reject H_0 if U_2 is too large. We calculate U_2 using the matrix $(\mathbf{Z}'\mathbf{Z})^{-1}$ computed in Example 11.5.3:

$$U_2 = \frac{0.6594 - 0.4}{(0.06762)^{1/2} \times 0.4011} = 2.487.$$

The degrees of freedom are $10 - 3 = 7$, and 2.487 falls between the 0.975 and 0.99 quantiles of the t distribution with seven degrees of freedom. The p-value is actually 0.0209, so we would reject H_0 at every level $\alpha_0 \geq 0.0209$. ◀

Problems of testing hypotheses that specify the values of two coefficients β_i and β_j are discussed in Exercises 17 to 21 at the end of this section. Problems of testing hypotheses about linear combinations of $\beta_0, \ldots, \beta_{p-1}$ are the subject of Exercise 26.

Some computer programs make it easy to test hypotheses about individual β_j's. Indeed, most software automatically supplies the value of the test statistic U_j for

testing the following hypotheses for each j $(j = 0, \ldots, k)$:

$$\begin{aligned} H_0: &\quad \beta_j = 0, \\ H_1: &\quad \beta_j \neq 0. \end{aligned} \tag{11.5.23}$$

Some programs also compute the corresponding p-values that are found from the expression (11.5.22).

Power of the Test If the null hypothesis in (11.5.20) is false, then the statistic U_j has the noncentral t distribution with $n - p$ degrees of freedom and noncentrality parameter $\psi = (\beta_j - \beta_j^*)/(\zeta_{jj}^{1/2}\sigma)$. Plots such as those in Figures 9.12 and 9.14 or computer programs can be used to calculate the power of the t test for specific parameter values.

Prediction

Let $\boldsymbol{z}' = (z_0, \ldots, z_{p-1})$ be a vector of predictors for a future observation Y. We wish to predict Y using $\hat{Y} = \boldsymbol{z}'\hat{\boldsymbol{\beta}}$, and we want to know the M.S.E. We shall assume that Y is independent of the observed data. This makes Y and $\hat{Y}$ independent. We can write

$$\hat{Y} = \boldsymbol{z}'\hat{\boldsymbol{\beta}} = \boldsymbol{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{Y},$$

so that $\hat{Y}$ is a linear combination of the original data $\boldsymbol{Y}$. Since the coordinates of $\boldsymbol{Y}$ are independent normal random variables, Theorem 11.3.1 tells us that $\hat{Y}$ has a normal distribution. The mean of $\hat{Y}$ is easily seen to be

$$E(\hat{Y}) = \boldsymbol{z}'E(\hat{\boldsymbol{\beta}}) = \boldsymbol{z}'\boldsymbol{\beta}.$$

The variance of $\hat{Y}$ is obtained from Theorem 11.5.2:

$$\begin{aligned} \text{Var}(\hat{Y}) &= \boldsymbol{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \, \text{Cov}(\boldsymbol{Y})\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\boldsymbol{z} \\ &= \boldsymbol{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\boldsymbol{z}\sigma^2. \end{aligned}$$

Since Y has the normal distribution with mean $\boldsymbol{z}'\boldsymbol{\beta}$ and variance σ^2 and is independent of $\hat{Y}$, it follows that $Y - \hat{Y}$ has the normal distribution with mean 0 and variance

$$\text{Var}(Y - \hat{Y}) = \text{Var}(\hat{Y}) + \text{Var}(Y) = \sigma^2\Big[1 + \boldsymbol{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\boldsymbol{z}\Big]. \tag{11.5.24}$$

Since $Y - \hat{Y}$ has mean 0, Eq. (11.5.24) is also the M.S.E. for using $\hat{Y}$ to predict Y.

We can also form a prediction interval for Y just as we did in (11.3.25). As we did there, define

$$Z = \frac{Y - \hat{Y}}{\sigma[1 + \boldsymbol{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\boldsymbol{z}]^{1/2}}, \quad W = \frac{S^2}{\sigma^2}.$$

Then Z has the standard normal distribution independent of W, which has the χ^2 distribution with $n - p$ degrees of freedom. Hence,

$$\frac{Z}{(W/[n-p])^{1/2}} = \frac{Y - \hat{Y}}{\sigma'[1 + \boldsymbol{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\boldsymbol{z}]^{1/2}}$$

has the t distribution with $n - p$ degrees of freedom. It follows that the interval with

the following endpoints has probability $1-\alpha_0$ of containing Y, prior to observing the data:

$$\hat{Y} \pm T^{-1}_{n-p}\left(1-\frac{\alpha_0}{2}\right)\sigma'\left[1+\boldsymbol{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\boldsymbol{z}\right]^{1/2}. \tag{11.5.25}$$

Example 11.5.7

Predicting Dishwasher Shipments. In Example 11.5.4, the least-squares estimates for the model are $\hat{\beta}_0 = -1314$, $\hat{\beta}_1 = 66.91$, and $\hat{\beta}_2 = 58.86$. The observed value of σ' is 352.9. Now suppose that we are interested in predicting dishwasher shipments for 1986. We happen to know that in 1986 private residential investment was 67.2 billion. In order to predict dishwasher shipments for 1986, we first form the vector of predictors $\boldsymbol{z}' = (1, 26, 67.2)$. Then we compute $\hat{Y} = \boldsymbol{z}'\hat{\boldsymbol{\beta}} = 4381$ and

$$\sigma'[1+\boldsymbol{z}'(\mathbf{Z}'\mathbf{Z})^{-1}\boldsymbol{z}]^{1/2} = 352.9[1+0.2136]^{1/2} = 388.8.$$

We can now compute a prediction interval for 1986 dishwasher shipments. For example, with $\alpha_0 = 0.1$, we get a 90 percent prediction interval using $T^{-1}_{23}(0.95) = 1.714$,

$$(4381 - 1.714 \times 388.8,\ 4381 + 1.714 \times 388.8) = (3715, 5047).$$

This is quite a wide range due to the large value of σ'. The actual value for dishwasher sales in 1986 was 3915, which is quite far from $\hat{Y}$, but still within the interval. ◀

Multiple R^2

In a problem of multiple linear regression, we are typically interested in determining how well the variables $X_1, \dots, X_k$ explain the observed variation in the random variable Y. The variation among the n observed values $y_1, \dots, y_n$ of Y can be measured by the value of $\sum_{i=1}^n (y_i - \overline{y})^2$, which is the sum of the squares of the deviations of $y_1, \dots, y_n$ from the average $\overline{y}$. Similarly, after the regression of Y on $X_1, \dots, X_k$ has been fitted from the data, the variation among the n observed values of Y that is still present can be measured by the sum of the squares of the deviations of $y_1, \dots, y_n$ from the fitted regression. This sum of squares will be equal to the value of S^2 in Eq. (11.5.6) calculated from the observed values, i.e., $S^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$.

It now follows that the proportion of the variation among the observed values $y_1, \dots, y_n$ that remains unexplained by the fitted regression is

$$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2}.$$

In turn, the proportion of the variation among the observed values $y_1, \dots, y_n$ that *is explained* by the fitted regression is given by the following value R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2}. \tag{11.5.26}$$

Example 11.5.8

Unemployment in the 1950s. For the data in Example 11.5.1, we can compute $\overline{y}_{10} = 2.82$, and then $\sum_{i=1}^{10} (y_i - \overline{y}_{10})^2 = 8.376$. The value of S^2 is $(10-3) \times \sigma'^2 = 1.126$, so $R^2 = 1 - 1.126/8.376 = 0.8656$. ◀

The value of R^2 must lie in the interval $0 \le R^2 \le 1$. When $R^2 = 0$, the least-squares estimates have the values $\hat{\beta}_0 = \overline{y}$ and $\hat{\beta}_1 = \cdots = \hat{\beta}_k = 0$. In this case, the fitted regression function is just the constant function $y = \overline{y}$. When R^2 is close to 1, the

variation of the observed values of Y around the fitted regression function is much smaller than their variation around $\overline{y}$.

Analysis of Residuals

In Sec. 11.3, we described some plots for assessing whether or not the assumptions of the simple linear regression model seem to be met. These same plots, together with some others, are also useful in the general linear model. Recall that, in general, the residuals are the values

$$e_i = y_i - \hat{y}_i = y_i - z_{i0}\beta_0 - \cdots - z_{ip-1}\beta_{p-1}.$$

Example 11.5.9 Unemployment in the 1950s. In this example, $p = 3$ with $z_{i0} = 1$ for all i. We have plotted the residuals against the two predictor variables in the top row of Fig. 11.17 to begin looking for violations of the assumptions. The residual from the first year (1950) is very high, and the remaining residuals appear to lie near a line with positive slope in each plot. This suggests that the first observation does not follow the same pattern as the others. We also performed the regression without the 1950 data point. The residual plots using the new least-squares estimates fit from the 1951–1959 data are in the bottom row of Fig. 11.17. The residuals for 1951–1959 no longer lie on a sloped line. Also, Fig. 11.18 shows normal quantile plots both before and after deleting the 1950 observation. The right plot is much straighter. Of course, such a graphical analysis does not show that the 1950 observation should be deleted. We should check to see if something might have occurred in 1950 that would make a drastic change to the relationship between unemployment and time (such as the start of the war in Korea.) ◄

Another plot that is useful in multiple regression cases is a plot of residuals against fitted values, $\hat{y}_i$ for $i = 1, \ldots, n$. (See Exercise 27 to see why this plot is not used in simple linear regression.) This plot helps to reveal dependence between the mean and variance of Y. (Recall that $\hat{y}_i$ is an estimate of the mean of Y_i.) If the residuals are more spread out at one end or the other of this plot, it suggests that the variance of Y changes as the mean changes, which violates the assumption that all observations have the same variance. The left plot in Fig. 11.19 is a plot of residuals

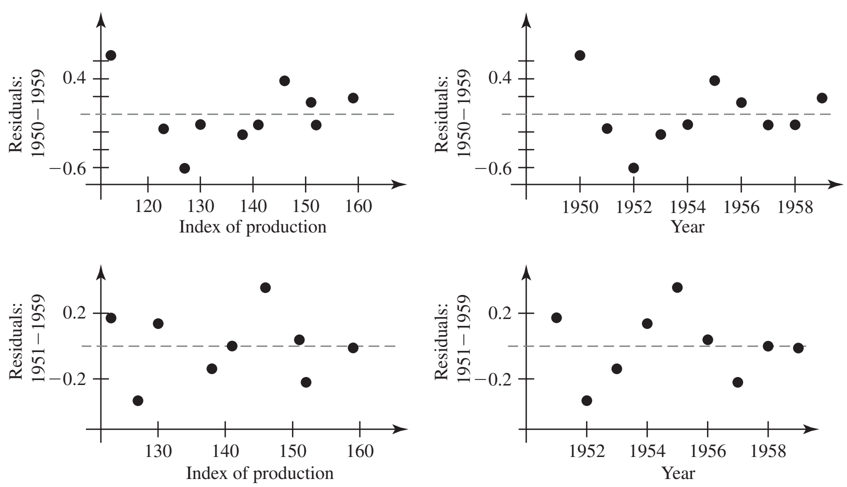


Figure 11.17 Plots of residuals against the two predictor variables for Example 11.5.9. Top row: using all data for 1950–1959. Bottom row: using only 1951–1959 data.

Figure 11.18 Normal quantile plots of residuals for Example 11.5.9. The left plot is from the regression using all 10 observations. The right plot uses only 1951–1959.

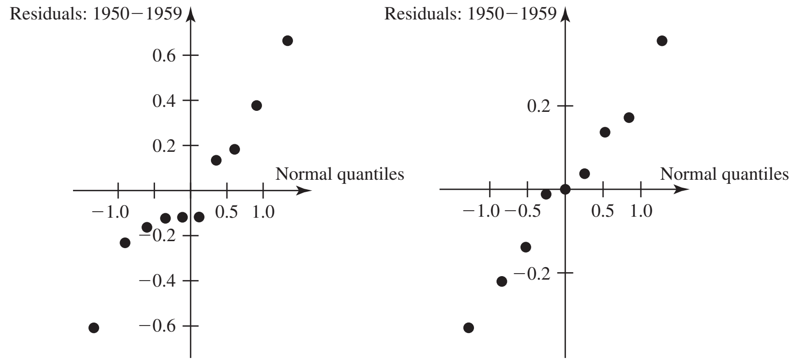
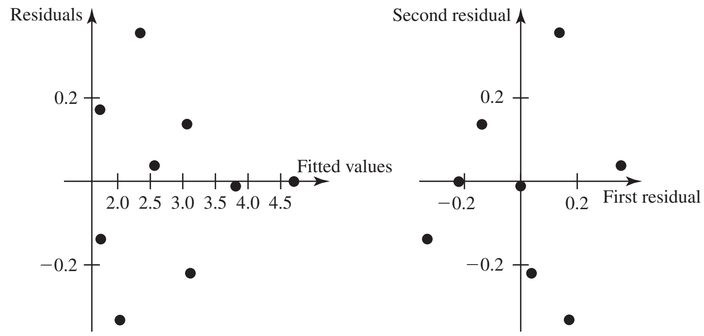


Figure 11.19 Residual plots for Example 11.5.9. Left: plot of residuals against fitted values. Right: plot of pairs of consecutive residuals. Both plots use 1951–1959 data only.



against fitted values for the unemployment data. It appears that the residuals corresponding to low fitted values are more spread out than those corresponding to high fitted values. Methods for responding to such features in a residual plot can be found in texts on regression methodology such as Draper and Smith (1998) and Cook and Weisberg (1999).

If the time of each measurement is available, as in Examples 11.5.1 and 11.5.4, it makes sense to plot residuals against time to see if there is any time dependence not captured by the model. Since time was one of the predictors in each of these examples, we will plot residuals against time when we plot residuals against the predictors. In addition to plotting the residuals against time, we can also plot the nearby residuals against each other to see if small ones tend to occur together and/or if large ones tend to occur together. Let $v_1, \ldots, v_n$ be the residuals ordered by time. We can plot the $n-1$ points $(v_1, v_2), (v_2, v_3), \ldots, (v_{n-1}, v_n)$. If these plotted points follow a pattern, it suggests that there is dependence between observations that are close together in time, called *serial dependence*. This would violate the assumption that the observations are independent. The right plot in Fig. 11.19 is the plot of consecutive pairs of residuals for the unemployment data. The points in this plot cluster in opposite corners, suggesting serial dependence, although the small sample size makes it difficult to be certain.

Example 11.5.10

Dishwasher Shipments. Consider, again, the data from Example 11.5.4. Plots of residuals against the two predictors, in the top row of Fig. 11.20, reveal a serious problem. There is a curve in the plot of residuals against the year. The residuals are highest in the middle years and lower in the early and late years. This suggests that perhaps the relationship between shipments and time is not linear. The plot of pairs of consecu-

Figure 11.20 Residual plots for Example 11.5.10. Top row: residuals against predictors. Lower left: residuals against fitted values. Lower right: pairs of successive residuals.

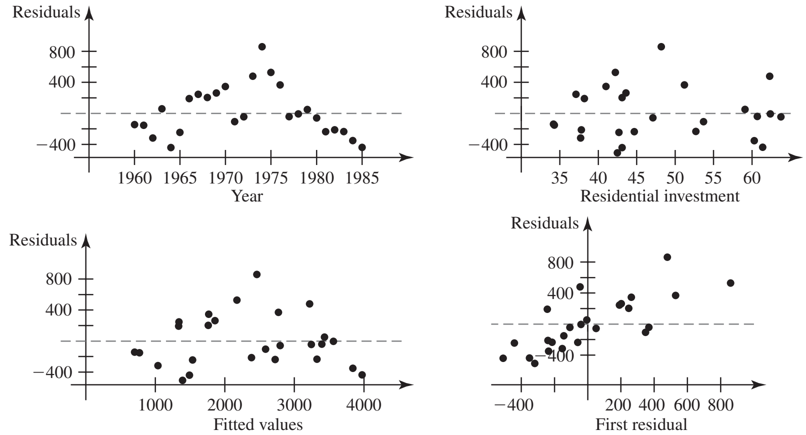
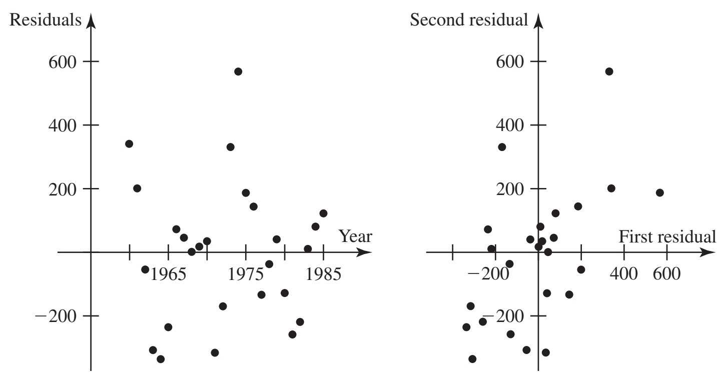


Figure 11.21 Residual plots for regression of dishwasher shipments on a quadratic function of time. Left: plot of residuals against time. Right: plot of pairs of consecutive residuals.



tive residuals also suggests some time dependence. This could be a result of the same problem that caused the curve in the plot of residuals against time, or it could indicate that successive observations are dependent. It is possible that deviations from the overall trend in dishwasher sales might persist for more than one year. For example, a boom or bust in sales one year might carry over to part of the next year. The normal quantile plot (not shown) is fairly straight.

In order to try to determine whether there is serial dependence or a nonlinear relationship (or both) in these data, we fit another model in which the mean of Y is a linear function of private residential investment but a quadratic function of time. That is, let X_1 stand for the year (minus 1960), let X_2 stand for private residential investment, and let $X_3 = X_1^2$. Then

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2.$$

The least-squares estimates from this model are $\hat{\beta}_0 = -1445$, $\hat{\beta}_1 = 206.1$, $\hat{\beta}_2 = 48.5$, and $\hat{\beta}_3 = -5.23$. The observed value of σ' is 235.7. The plots of residuals against time and of consecutive pairs of residuals are in Fig. 11.21. The plot of residuals against time is better than before, but the pairs of consecutive residuals still lie close to a line. This suggests that we need to take serial dependence into account. One book that describes methods for dealing with serial dependence (commonly called *time series analysis*) is Box, Jenkins, and Reinsel (1994). ◀

Summary

In the general linear model, we assume that the mean of each observation Y_i can be expressed as $z_{i0}\hat{\beta}_0 + \cdots + z_{ip-1}\hat{\beta}_{p-1}$, where $\beta_0, \ldots, \beta_{p-1}$ are unknown parameters and $z_{i0}, \ldots, z_{ip-1}$ are the observed values of predictors. These predictors can be control variables, other variables that are measured along with Y_i, or functions of such variables. Least-squares estimators of the parameters are denoted $\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}$, and they can be calculated according to Eq. (11.5.10) or by using a computer. The variance of each Y_i is assumed to be the same value σ^2. Every linear combination of the least-squares estimators has a normal distribution and is independent of the unbiased estimator σ'^2 of σ^2 given in Eq. (11.5.8).

For testing hypotheses about a single β_j, the statistic U_j in Eq. (11.5.21) has the t distribution with $n - p$ degrees of freedom given that the null hypothesis is true. For predicting a future Y value, we can form prediction intervals using the endpoints given by (11.5.25). We should always plot the residuals $y_i - \hat{y}_i$ against the predictors, fitted values $\hat{y}_i$, and time (if available) to check on the assumptions of the linear regression model. Patterns in these plots can suggest violations of the assumption about the form of the mean of Y_i and/or the constant variance assumption. We should also make a normal quantile plot. Deviations from a straight line in this plot suggest that the Y_i values might not have a normal distribution, although violations of the assumptions about the mean and variance can also cause patterns in this plot. If observation time is available, we should also plot pairs of consecutive residuals to look for serial dependence.

Exercises

1. Show that the M.L.E. of σ^2 in the general linear model is given by Eq. (11.5.7).

2. Prove that σ'^2, defined in Eq. (11.5.8), is an unbiased estimator of σ^2. You may assume that S^2 has a χ^2 distribution with $n - p$ degrees of freedom.

3. Consider a regression problem in which, for each value x of a certain variable X, the random variable Y has the normal distribution with mean βx and variance σ^2, where the values of β and σ^2 are unknown. Suppose that n independent pairs of observations (x_i, Y_i) are obtained. Show that the M.L.E. of β is

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}.$$

4. For the conditions of Exercise 3, show that $E(\hat{\beta}) = \beta$ and $\text{Var}(\hat{\beta}) = \sigma^2 / (\sum_{i=1}^{n} x_i^2)$.

5. Suppose that when a small amount x of an insulin preparation is injected into a rabbit, the percentage decrease Y in blood sugar has the normal distribution with mean βx and variance σ^2, where the values of β and σ^2 are unknown. Suppose that when independent observations are made on 10 different rabbits, the observed values of x_i and Y_i for $i = 1, \ldots, 10$ are as given in Table 11.14. Determine the values of the M.L.E.'s $\hat{\beta}$ and $\hat{\sigma}^2$, and the value of $\text{Var}(\hat{\beta})$.

Table 11.14 Data for Exercise 5

i	x_i	y_i	i	x_i	y_i
1	0.6	8	6	2.2	19
2	1.0	3	7	2.8	9
3	1.7	5	8	3.5	14
4	1.7	11	9	3.5	22
5	2.2	10	10	4.2	22

6. For the conditions of Exercise 5 and the data in Table 11.14, carry out a test of the following hypotheses:

$$H_0: \quad \beta = 10,$$
$$H_1: \quad \beta \neq 10.$$

7. Consider a regression problem in which a patient's reaction Y to a new drug B is to be related to his reaction X to a standard drug A. Suppose that for each value x of X, the regression function is a polynomial of the form $E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$. Suppose also that 10 pairs of observed values are as shown in Table 11.1 on page 690. Under the standard assumptions of the general linear model, determine the values of the M.L.E.'s $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$.

8. For the conditions of Exercise 7 and the data in Table 11.1, determine the values of $\text{Var}(\hat{\beta}_0)$, $\text{Var}(\hat{\beta}_1)$, $\text{Var}(\hat{\beta}_2)$, $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$, $\text{Cov}(\hat{\beta}_0, \hat{\beta}_2)$, and $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$.

9. For the conditions of Exercise 7 and the data in Table 11.1, carry out a test of the following hypotheses:

$$\begin{aligned} H_0&: \quad \beta_2 = 0,\\ H_1&: \quad \beta_2 \neq 0. \end{aligned}$$

10. For the conditions of Exercise 7 and the data in Table 11.1, carry out a test of the following hypotheses:

$$\begin{aligned} H_0&: \quad \beta_1 = 4,\\ H_1&: \quad \beta_1 \neq 4. \end{aligned}$$

11. For the conditions of Exercise 7 and the data given in Table 11.1, determine the value of R^2, as defined by Eq. (11.5.26).

12. Consider a problem of multiple linear regression in which a patient's reaction Y to a new drug B is to be related to her reaction X_1 to a standard drug A and her heart rate X_2. Suppose that, for all values $X_1 = x_1$ and $X_2 = x_2$, the regression function has the form $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, and the values of 10 sets of observations (x_{i1}, x_{i2}, Y_i) are given in Table 11.2 on page 696. Under the standard assumptions of multiple linear regression, determine the values of the M.L.E.'s $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$.

13. For the conditions of Exercise 12 and the data in Table 11.2, determine the values of $\text{Var}(\hat{\beta}_0)$, $\text{Var}(\hat{\beta}_1)$, $\text{Var}(\hat{\beta}_2)$, $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$, $\text{Cov}(\hat{\beta}_0, \hat{\beta}_2)$, and $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$.

14. For the conditions of Exercise 12 and the data in Table 11.2, carry out a test of the following hypotheses:

$$\begin{aligned} H_0&: \quad \beta_1 = 0,\\ H_1&: \quad \beta_1 \neq 0. \end{aligned}$$

15. For the conditions of Exercise 12 and the data in Table 11.2, carry out a test of the following hypotheses:

$$\begin{aligned} H_0&: \quad \beta_2 = -1,\\ H_1&: \quad \beta_2 \neq -1. \end{aligned}$$

16. For the conditions of Exercise 12 and the data in Table 11.2, determine the value of R^2, as defined by Eq. (11.5.26).

17. Consider the general linear model in which the observations $Y_1, \dots, Y_n$ are independent and have normal distributions with the same variance σ^2 and in which $E(Y_i)$ is given by Eq. (11.5.1). Let the matrix $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ be defined by Eq. (11.5.19). For all values of i and j such that $i \neq j$, let the random variable A_{ij} be defined as follows:

$$A_{ij} = \hat{\beta}_i - \frac{\zeta_{ij}}{\zeta_{jj}} \hat{\beta}_j.$$

Show that $\text{Cov}(\hat{\beta}_j, A_{ij}) = 0$, and explain why $\hat{\beta}_j$ and A_{ij} are therefore independent.

18. For the conditions of Exercise 17, show that $\text{Var}(A_{ij}) = [\zeta_{ii} - (\zeta_{ij}^2/\zeta_{jj})]\sigma^2$. Also show that the following random variable W^2 has the χ^2 distribution with two degrees of freedom:

$$W^2 = \frac{\zeta_{jj}(\hat{\beta}_i - \beta_i)^2 + \zeta_{ii}(\hat{\beta}_j - \beta_j)^2 - 2\zeta_{ij}(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j)}{\left(\zeta_{ii}\zeta_{jj} - \zeta_{ij}^2\right)\sigma^2}.$$

Hint: Show that

$$W^2 = \frac{(\hat{\beta}_j - \beta_j)^2}{\zeta_{jj}\sigma^2} + \frac{\left[A_{ij} - E(A_{ij})\right]^2}{\text{Var}(A_{ij})}.$$

19. Consider again the conditions of Exercises 17 and 18, and let the random variable σ' be as defined by Eq. (11.5.8).

a. Show that the random variable $\sigma^2 W^2/(2\sigma'^2)$ has the F distribution with two and $n - p$ degrees of freedom.

b. For every two given numbers β_i^* and β_j^*, describe how to carry out a test of the following hypotheses:

$$\begin{aligned} H_0&: \quad \beta_i = \beta_i^* \text{ and } \beta_j = \beta_j^*,\\ H_1&: \quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned}$$

20. For the conditions of Exercise 7 and the data in Table 11.1, carry out a test of the following hypotheses:

$$\begin{aligned} H_0&: \quad \beta_1 = \beta_2 = 0,\\ H_1&: \quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned}$$

21. For the conditions of Exercise 12 and the data in Table 11.2, carry out a test of the following hypotheses:

$$\begin{aligned} H_0&: \quad \beta_1 = 1 \text{ and } \beta_2 = 0,\\ H_1&: \quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned}$$

22. Consider a problem of simple linear regression as described in Sec. 11.2, and let R^2 be defined by Eq. (11.5.26) of this section. Show that

$$R^2 = \frac{\left[\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})\right]^2}{\left[\sum_{i=1}^n (x_i - \overline{x})^2\right]\left[\sum_{i=1}^n (y_i - \overline{y})^2\right]}.$$

23. Suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are n-dimensional random vectors for which the mean vectors $E(\boldsymbol{X})$ and $E(\boldsymbol{Y})$ exist. Show that $E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y})$.

24. Suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent n-dimensional random vectors for which the covariance matrices $\text{Cov}(\boldsymbol{X})$ and $\text{Cov}(\boldsymbol{Y})$ exist. Show that $\text{Cov}(\boldsymbol{X} + \boldsymbol{Y}) = \text{Cov}(\boldsymbol{X}) + \text{Cov}(\boldsymbol{Y})$.

25. Suppose that $\boldsymbol{Y}$ is a three-dimensional random vector with coordinates Y_1, Y_2, and Y_3, and suppose that the covariance matrix of $\boldsymbol{Y}$ is as follows:

$$\text{Cov}(\boldsymbol{Y}) = \begin{bmatrix} 9 & -3 & 0 \\ -3 & 4 & 0 \\ 0 & 0 & 5 \end{bmatrix}.$$

Determine the value of $\text{Var}(3Y_1 + Y_2 - 2Y_3 + 8)$.

26. In a general linear model setting with p predictors, we wish to test the following hypotheses:

$$\begin{aligned} H_0: \quad & \sum_{j=0}^{p-1} c_j \beta_j = c_*, \\ H_1: \quad & \sum_{j=0}^{p-1} c_j \beta_j \neq c_*. \end{aligned} \tag{11.5.27}$$

a. Show that $\sum_{j=0}^{p-1} c_j \hat{\beta}_j$ has a normal distribution and find its mean and variance. (You may wish to use Theorems 11.3.1 and 11.5.2.)

b. Let $\boldsymbol{c}' = (c_0, \ldots, c_{p-1})$. If H_0 is true, show that

$$U = \frac{\sum_{j=0}^{p-1} c_j \hat{\beta}_j - c_*}{\sigma' (\boldsymbol{c}'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{c})^{1/2}}$$

has the t distribution with $n - p$ degrees of freedom.

c. Explain how to test the hypotheses in (11.5.27) at level of significance α_0.

27. In a simple linear regression problem, the plot of residuals against fitted values would look the same as the plot of residuals against the predictor X (or a mirror image of it), except for the labeling of the horizontal axis. Explain why this is true.

28. Consider a multiple linear regression problem with design matrix $\boldsymbol{Z}$ and observations $\boldsymbol{Y}$. Let $\boldsymbol{Z}_1$ be the matrix remaining when at least one column is removed from $\boldsymbol{Z}$. Then $\boldsymbol{Z}_1$ is the design matrix for a linear regression problem with fewer predictors and the same data $\boldsymbol{Y}$. Prove that the value of R^2 calculated in the problem using design matrix $\boldsymbol{Z}$ is at least as large as the value of R^2 calculated in the problem using design matrix $\boldsymbol{Z}_1$.

29. Calculate the value of R^2 for the dishwasher shipment data (Example 11.5.4) using the model in which the mean of Y_i is a linear function of both year and private residential investment.

30. Consider again the conditions of Exercise 26. Suppose that the null hypothesis in (11.5.27) is false. Find the distribution of the statistic U defined in that exercise.

11.6 Analysis of Variance

In Sec. 9.6, we studied methods for comparing the means of two normal distributions. In this section, we shall consider experiments in which we need to compare the means of two or more normal distributions. The theory behind the methods developed here is based entirely on results from the general linear model in Sec. 11.5.

The One-Way Layout

Example 11.6.1 Calories in Hot Dogs. Moore and McCabe (1999) describe data gathered by *Consumer Reports* (June 1986, pp. 364–67). The data comprise (among other things) calorie contents from 63 brands of hot dogs. (See Table 11.15.) The hot dogs come in four varieties: beef, "meat" (don't ask), poultry, and "specialty." (Specialty hot dogs include stuffing such as cheese or chili.) It is interesting to know whether, and to what extent, the different varieties differ in their calorie contents. Data structures of the sort in this example, consisting of several groups of similar random variables, are the subject of this section. ◀

In this section and in the remainder of this chapter, we shall study a topic known as the *analysis of variance*, abbreviated ANOVA. Problems of ANOVA are actually problems of multiple regression in which the design matrix $\boldsymbol{Z}$ has a very special form. In other words, the study of ANOVA can be placed within the framework of the general linear model (Definition 11.5.1), if we continue to make

Table 11.15 Calorie counts in four types of hot dogs for Example 11.6.2

Type	Calorie Count
Beef	186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111, 141, 153, 190, 157, 131, 149, 135, 132
Meat	173, 191, 182, 190, 172, 147, 146, 139, 175, 136, 179, 153, 107, 195, 135, 140, 138
Poultry	129, 132, 102, 106, 94, 102, 87, 99, 107, 113, 135, 142, 86, 143, 152, 146, 144
Specialty	155, 170, 114, 191, 162, 146, 140, 187, 180

the basic assumptions for such a model: The observations that are obtained are independent and normally distributed; all these observations have the same variance σ^2; and the mean of each observation can be represented as a linear combination of certain unknown parameters. The theory and methodology of ANOVA were mainly developed by R. A. Fisher during the 1920s.

We shall begin our study of ANOVA by considering a problem known as the *one-way layout*. In this problem, it is assumed that random samples from p different normal distributions are available, each of these distributions has the same variance σ^2, and the means of the p distributions are to be compared on the basis of the observed values in the samples. This problem was considered for two populations $(p = 2)$ in Sec. 9.6, and the results to be presented here for an arbitrary value of p will generalize those presented in Sec. 9.6. Specifically, we shall now make the following assumption: For $i = 1, \dots, p$, the random variables $Y_{i1}, \dots, Y_{in_i}$, form a random sample of n_i observations from the normal distribution with mean μ_i and variance σ^2, and the values of $\mu_1, \dots, \mu_p$ and σ^2 are unknown.

In this problem, the sample sizes $n_1, \dots, n_p$ are not necessarily the same. We shall let $n = \sum_{i=1}^{p} n_i$ denote the total number of observations in the p samples, and we shall assume that all n observations are independent.

Example 11.6.2 Calories in Hot Dogs. In Example 11.6.1, the sample sizes are $n_1 = 20$ (beef), $n_2 = 17$ (meat), $n_3 = 17$ (poultry), and $n_4 = 9$ (specialty). In this case, we let μ_1 stand for the mean calorie count for brands of beef hot dogs, while μ_2, μ_3, and μ_4 will stand for the mean calorie count for brands of meat, poultry, and specialty hot dogs, respectively. All calorie counts are assumed to be independent normal random variables with variance σ^2. These data will be analyzed after we develop the ANOVA methodology. ◄

It follows from the assumptions we have just made that for $j = 1, \dots, n_i$ and $i = 1, \dots, p$, we have $E(Y_{ij}) = \mu_i$ and $\text{Var}(Y_{ij}) = \sigma^2$. Since the expectation $E(Y_{ij})$ of each observation is equal to one of the p parameters $\mu_1, \dots, \mu_p$, it is obvious that each of these expectations can be regarded as a linear combination of $\mu_1, \dots, \mu_p$. Furthermore, we can regard the n observations Y_{ij} as the elements of a single long

n-dimensional vector $\boldsymbol{Y}$, which can be written as follows:

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{p1} \\ \vdots \\ Y_{pn_p} \end{bmatrix}. \tag{11.6.1}$$

This one-way layout therefore satisfies the conditions of the general linear model. In order to make the one-way layout look exactly like the general linear model, we could define parameters $\beta_i = \mu_{i+1}$ for $i = 0, \ldots, p-1$. Then the $n \times p$ design matrix, $\boldsymbol{Z}$, has one column for each population. The column corresponding to population 1 has n_1 1's followed by $n_2 + \cdots + n_p$ 0's. The column corresponding to population 2 has n_1 0's followed by n_2 1's followed by $n_3 + \cdots + n_p$ 0's, and so on. For example, using the hot dog data in Example 11.6.2, the $\boldsymbol{Z}$ matrix would be

$$\boldsymbol{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ & \vdots & & \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ & \vdots & & \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ & \vdots & & \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ & \vdots & & \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{20 rows} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{17 rows} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{17 rows} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{9 rows} \end{matrix} \tag{11.6.2}$$

We shall not use the general linear model notation any further in the development of ANOVA, because the parameters $\mu_1, \ldots, \mu_p$ are more natural.

For $i = 1, \ldots, p$, we shall let $\overline{Y}_{i+}$ denote the sample mean of the n_i observations in the ith sample. Thus,

$$\overline{Y}_{i+} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}. \tag{11.6.3}$$

Similar logic to that used in the proof of Theorem 11.2.1 can be used to show that $\overline{Y}_{i+}$ is the M.L.E., or least-squares estimator, of μ_i for $i = 1, \ldots, p$. Also, the M.L.E. of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i+}\right)^2. \tag{11.6.4}$$

The details are left to Exercise 1.

Partitioning a Sum of Squares

Example 11.6.3 Calories in Hot Dogs. In Examples 11.6.1 and 11.6.2, we notice that the calorie counts within each type differ quite a bit from each other. We need to be able to quantify both the variation within type and the variation between types if we are going to try to address the question of whether or not different types of hot dogs have the same calorie counts. ◀

In a one-way layout, we are often interested in testing the hypothesis that the p distributions from which the samples were drawn are actually the same; that is, we desire to test the following hypotheses:

$$\begin{aligned} H_0: &\quad \mu_1 = \cdots = \mu_p, \\ H_1: &\quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{11.6.5}$$

For instance, in Example 11.6.2, the null hypothesis H_0 in (11.6.5) would be that the mean calorie counts for all four types of hot dogs are the same, but it would not specify what the common value is. The alternative hypothesis H_1 would be that at least two of the means differ, but it would not specify which means differ nor would it specify by how much the means differ.

Before we develop an appropriate test procedure, we shall carry out some preparatory algebraic manipulations. First, define

$$\overline{Y}_{++} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n}\sum_{i=1}^{p} n_i \overline{Y}_{i+},$$

which is the overall average of all n observations. We shall partition the sum of squares

$$S^2_{\text{Tot}} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{++})^2 \tag{11.6.6}$$

into two smaller sums of squares, each of which will be associated with a certain type of variation among the n observations. Note that S^2_{Tot}/n would be the M.L.E. of σ^2 if we believed that all of the observations came from a single normal distribution rather than from p different normal distributions. This means that we can interpret S^2_{Tot} as an overall measure of variation between the n observations. One of the smaller sums of squares into which we shall partition S^2_{Tot} will measure the variation between the p different samples, and the other sum of squares will measure the variation between the observations within each of the samples. The test of the hypotheses (11.6.5) that we shall develop will be based on the ratio of these two measures of variation. For this reason, the name *analysis of variance* has been used to describe this problem and other related problems.

Theorem 11.6.1 Partitioning the Sum of Squares. Let S^2_{Tot} be as defined in Eq. (11.6.6). Then

$$S^2_{\text{Tot}} = S^2_{\text{Resid}} + S^2_{\text{Betw}}, \tag{11.6.7}$$

where

$$S^2_{\text{Resid}} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i+})^2, \quad \text{and} \quad S^2_{\text{Betw}} = \sum_{i=1}^{p} n_i(\overline{Y}_{i+} - \overline{Y}_{++})^2.$$

Furthermore, $S^2_{\text{Resid}}/\sigma^2$ has the χ^2 distribution with $n - p$ degrees of freedom and is independent of S^2_{Betw}.

Table 11.16 General form of ANOVA table for one-way layout

Source of variation	Degrees of freedom	Sum of squares	Mean square
Between samples	$p-1$	S^2_{Betw}	$S^2_{\text{Betw}}/(p-1)$
Residuals	$n-p$	S^2_{Resid}	$S^2_{\text{Resid}}/(n-p)$
Total	$n-1$	S^2_{Tot}	

Proof If we consider only the n_i observations in sample i, then the sum of squares for those values can be written as follows:

$$\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{++})^2=\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i+})^2+n_i(\overline{Y}_{i+}-\overline{Y}_{++})^2. \tag{11.6.8}$$

It follows from Theorem 8.3.1 that the sum forming the first term on the right side of Eq. (11.6.8) has the χ^2 distribution with n_i-1 degrees of freedom and that it is independent of $\overline{Y}_{i+}$. Since $\overline{Y}_{++}$ is a function of $\overline{Y}_{1+},\ldots,\overline{Y}_{p+}$, all of which are independent of the first term on the right side of Eq. (11.6.8), it follows that the two terms on the right side of Eq. (11.6.8) are independent.

If we now sum each of the terms in Eq. (11.6.8) over the values of i, we obtain Eq. (11.6.7). Since all the observations in the p samples are independent, the two terms on the right side of Eq. (11.6.7) are independent. Also, $S^2_{\text{Resid}}/\sigma^2$ is the sum of p independent random variables, with the ith one having the χ^2 distribution with n_i-1 degrees of freedom. Hence, $S^2_{\text{Resid}}/\sigma^2$ will itself have the χ^2 distribution with $\sum_{i=1}^{p}(n_i-1)=n-p$ degrees of freedom. ■

As we noted earlier, S^2_{Tot} can be regarded as the total variation of the observations around their overall mean. Similarly, S^2_{Resid} can be regarded as the total variation of the observations around their particular sample means, or the total residual variation within the samples. Also, S^2_{Betw} can be regarded as the total variation of the sample means around the overall mean, or the variation between the sample means. Thus, the total variation S^2_{Tot} has been partitioned into two independent components, S^2_{Resid} and S^2_{Betw}, which represent different types of variations. This partitioning is often summarized in a table, which is called the ANOVA table for the one-way layout and is presented here as Table 11.16.

The numbers in the "Mean square" column of Table 11.16 are just the sums of squares divided by the degrees of freedom. They are used for testing the hypotheses (11.6.5). The degrees of freedom in the "Between samples" and "Total" rows will turn out to be degrees of freedom for random variables with χ^2 distributions if the null hypothesis in (11.6.5) is true. We shall see why this is true after we develop an appropriate test of the hypotheses (11.6.5).

Note: The Residual Mean Square Is the Same as the Unbiased Estimator of σ^2 in the Regression Setting. We began this section by expressing the one-way layout as a multiple linear regression problem with data vector $\boldsymbol{Y}$ and design matrix $\boldsymbol{Z}$. Compare the M.L.E. of σ^2, $\hat{\sigma}^2$ in Eq. (11.6.4), to the residual mean square in Table 11.16 to see that the two differ only in the constant in the denominator. The M.L.E. is S^2_{Resid}/n,

Table 11.17 ANOVA table for Example 11.6.4

Source of variation	Degrees of freedom	Sum of squares	Mean square
Between samples	3	19,454	6485
Residuals	59	32,995	559.2
Total	62	52,449	

while the residual mean square is $S^2_{\text{Resid}}/(n-p)$. Recall that this last ratio was called σ'^2 in Sec. 11.5, and is an unbiased estimator of σ^2. (Prove this last fact in Exercise 8.)

Example 11.6.4 Calories in Hot Dogs. The four sample averages in Example 11.6.2 are

$$\overline{Y}_{1+} = 156.85, \quad \overline{Y}_{2+} = 158.71, \quad \overline{Y}_{3+} = 118.76, \quad \overline{Y}_{4+} = 160.56.$$

The overall average is $\overline{Y}_{++} = 147.60$. We can now form the ANOVA table in Table 11.17. We shall test the hypotheses (11.6.5) after we develop an appropriate test statistic. ◄

Testing Hypotheses

In order to test the hypotheses (11.6.5), we need a test statistic that will tend to be larger if H_1 is true than it is if H_0 is true. We also need to know the distribution of the test statistic when H_0 is true.

Theorem 11.6.2 Suppose that H_0 in (11.6.5) is true. Then

$$U^2 = \frac{S^2_{\text{Betw}}/(p-1)}{S^2_{\text{Resid}}/(n-p)} \tag{11.6.9}$$

has the F distribution with $p-1$ and $n-p$ degrees of freedom.

Proof If all p samples of observations have the same mean, it can be shown (see Exercise 2) that $S^2_{\text{Betw}}/\sigma^2$ has the χ^2 distribution with $p-1$ degrees of freedom. We have already seen that S^2_{Betw} is independent of S^2_{Resid}, and $S^2_{\text{Resid}}/\sigma^2$ has the χ^2 distribution with $n-p$ degrees of freedom. It therefore follows that when H_0 is true, U^2 has the distribution stated in the theorem. ■

When the null hypothesis H_0 is not true, so that at least two of the μ_i values are different, then the expectation of the numerator of U^2 will be larger than it would be if H_0 were true. (See Exercise 11.) The distribution of the denominator of U^2 remains the same regardless of whether or not H_0 is true. A sensible level α_0 test of the hypotheses (11.6.5) would then be to reject H_0 if $U^2 \geq F^{-1}_{p-1,\,n-p}(1-\alpha_0)$, where $F^{-1}_{p-1,\,n-p}$ is the quantile function for the F distribution with $p-1$ and $n-p$ degrees of freedom. A partial table of F distribution quantiles is given in the back of this book. It can be shown that this test is also the level α_0 likelihood ratio test procedure. (See Exercise 12.)

Example 11.6.5

Calories in Hot Dogs. Suppose that we desire to test the null hypothesis that all four types of hot dogs have the same mean calorie count against the alternative hypothesis that at least two types have different means. The statistic U^2 in Eq. (11.6.9) has the F distribution with 3 and 59 degrees of freedom if the null hypothesis is true. The observed value of U^2 is the ratio of the between samples mean square to the residual mean square from Table 11.17, namely, $6485/559.2 = 11.60$. The p-value corresponding to this value is 4.5×10^{-6}, so the null hypothesis would be rejected at most standard levels. ◀

Power of the Test If the null hypothesis in (11.6.5) is false, then the statistic U^2 in Eq. (11.6.9) has a distribution known as noncentral F. For more details on the power function, consult a more advanced text such as Scheffé (1959, chapter 2). We shall not discuss the power of ANOVA tests any further.

Analysis of Residuals

Since the one-way layout is a special case of the general linear model, we make the assumptions of the general linear model when we perform the one-way ANOVA calculations. We should also compute residuals and plot them to see if the assumptions appear reasonable. The residuals are the values $e_{ij} = Y_{ij} - \bar{Y}_{i+}$, for $j = 1, \dots, n_i$ and $i = 1, \dots, p$.

Example 11.6.6

Calories in Hot Dogs. Figure 11.22 contains a plot of residuals against the categorical variable "hot dog type." Figure 11.23 contains the plot of residuals against normal quantiles. The points in the normal quantile plot are labeled by the hot dog type. Several disturbing features appear in these plots. First, there are three residuals with large negative values. Second, each of the first three samples appears to contain two distinct subsets, one with low residuals and one with high residuals. There is a gap between the two subsets in each sample. This suggests that there is another variable that we haven't discussed yet but which distinguishes these two subgroups. If we go back to the reported data (in the original *Consumer Reports* article), we find that the weight of each package and the number of hot dogs per package are also reported. The ratio of these two numbers is the weight of an average hot dog. Figure 11.24 contains a plot of residuals against average hot dog weight. Notice that most of the large residuals come from the larger (heavier) hot dogs and the smaller residuals tend to come from the smaller (lighter) hot dogs. Perhaps a better analysis would have set Y equal to calories per ounce rather than calories per hot dog. ◀

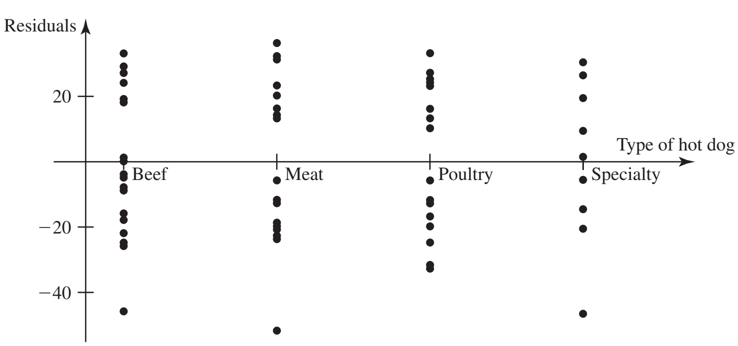


Figure 11.22 Plot of residuals against hot dog type.

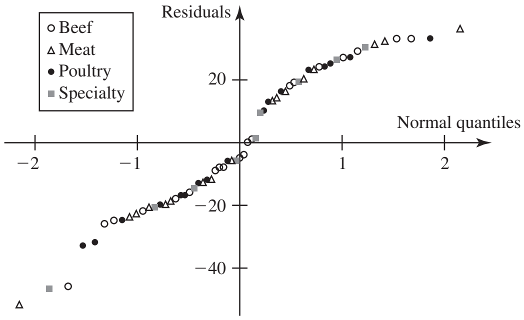


Figure 11.23 Plot of residuals against normal quantiles. The points are labeled by the hot dog type.

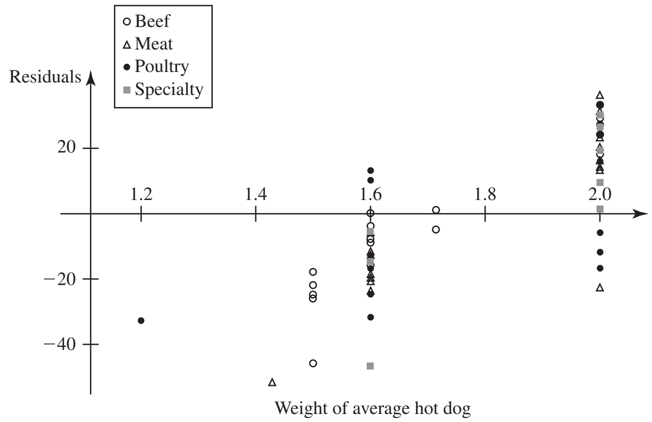


Figure 11.24 Plot of residuals against average hot dog weight. The points are labeled by the hot dog type.

Summary

The one-way layout can be considered as a general linear model, and we can use the methods of Sec. 11.5 to fit the model. However, the hypotheses of most interest in the one-way layout are (11.6.5). These hypotheses concern more than one linear combination of the regression coefficients, and they are not a special case of the hypotheses that we learned how to test in Sec. 11.5. To test these new hypotheses, we developed the analysis of variance (ANOVA) and the ANOVA table. The test statistic is U^2 in Eq. (11.6.9), which has the F distribution with $p-1$ and $n-p$ degrees of freedom if H_0 is true. The level α_0 test of H_0 is to reject H_0 if U^2 is greater than the $1-\alpha_0$ quantile of the appropriate F distribution.

Exercises

1. In a one-way layout, show that $\overline{Y}_{i+}$ is the least-squares estimator of μ_i by showing that the ith coordinate of the vector $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$ is $\overline{Y}_{i+}$ for $i=1,\dots,p$.

2. Assume that H_0 in (11.6.5) is true; that is, all observations have the same mean μ. Prove that $S^2_{\text{Betw}}/\sigma^2$ has the χ^2 distribution with $p-1$ degrees of freedom. *Hint:* Let

$$\boldsymbol{X}=\begin{pmatrix} n_1^{1/2}(\overline{Y}_{1+}-\mu)/\sigma^2 \\ \vdots \\ n_p^{1/2}(\overline{Y}_{p+}-\mu)/\sigma^2 \end{pmatrix},$$

then use the same method that was used in Sec. 8.3 to find the distribution of the sample variance. You may use the following fact without proving it:

Let $\boldsymbol{u}=((n_1/n)^{1/2},\dots,(n_p/n)^{1/2})$. *Then there exists an orthogonal matrix* $\boldsymbol{A}$ *whose first row is* $\boldsymbol{u}$.

3. Show that

$$\sum_{i=1}^{p} n_i(\overline{Y}_{i+} - \overline{Y}_{++})^2 = \sum_{i=1}^{p} n_i \overline{Y}_{i+}^2 - n\overline{Y}_{++}^2.$$

4. Specimens of milk from a number of dairies in three different districts were analyzed, and the concentration of the radioactive isotope strontium-90 was measured in each specimen. Suppose that specimens were obtained from four dairies in the first district, from six dairies in the second district, and from three dairies in the third district, and that the results measured in picocuries per liter were as follows:

District 1: 6.4, 5.8, 6.5, 7.7,

District 2: 7.1, 9.9, 11.2, 10.5, 6.5, 8.8,

District 3: 9.5, 9.0, 12.1.

a. Assuming that the variance of the concentration of strontium-90 is the same for the dairies in all three districts, determine the M.L.E. of the mean concentration in each of the districts and the M.L.E. of the common variance.

b. Test the hypothesis that the three districts have identical concentrations of strontium-90.

5. A random sample of 10 students was selected from the senior class at each of four large high schools, and the score of each of these 40 students on a certain mathematics examination was observed. Suppose that for the 10 students from each school, the sample mean and the sample variance of the scores were as shown in Table 11.18. Test the hypothesis that the senior classes at all four high schools would perform equally well on this examination. Discuss carefully the assumptions that you are making in carrying out this test.

Table 11.18 Data for Exercise 5

School	Sample mean	Sample variance
1	105.7	30.3
2	102.0	54.4
3	93.5	25.0
4	110.8	36.4

6. Suppose that a random sample of size n is taken from the normal distribution with mean μ and variance σ^2. Before the sample is observed, the random variables are divided into p groups of sizes $n_1, \dots, n_p$, where $n_i \geq 2$ for $i = 1, \dots, p$ and $\sum_{i=1}^{p} n_i = n$. For $i = 1, \dots, p$, let Q_i denote the sum of the squares of the deviations of the n_i observations in the ith group from the sample mean of those n_i observations. Find the distribution of the sum $Q_1 + \cdots + Q_p$ and the distribution of the ratio Q_1/Q_p.

7. Verify that the t test presented in Sec. 9.6 for comparing the means of two normal distributions is the same as the test presented in this section for the one-way layout with $p = 2$ by verifying that if U is defined by Eq. (9.6.3), then U^2 is equal to the expression given in Eq. (11.6.9).

8. Show that in a one-way layout the following statistic is an unbiased estimator of σ^2:

$$\frac{1}{n-p} \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i+})^2.$$

9. In a one-way layout, show that for all values of i, i', and j, where $j = 1, \dots, n_i$, $i = 1, \dots, p$, and $i' = 1, \dots, p$, the following three random variables W_1, W_2, and W_3 are uncorrelated with each other:

$$W_1 = Y_{ij} - \overline{Y}_{i+}, \quad W_2 = \overline{Y}_{i'+} - \overline{Y}_{++}, \quad W_3 = \overline{Y}_{++}.$$

10. In 1973, the President of Texaco, Inc., made a statement to a U.S. Senate subcommittee concerned with air and water pollution. The committee was concerned with, among other things, the noise levels associated with automobile filters. He cited the data in Table 11.19 from a study that included vehicles of three different sizes.

Table 11.19 Data for Exercise 10

Vehicle size	Noise values
Small	810, 820, 820, 835, 835, 835
Medium	840, 840, 840, 845, 855, 850
Large	785, 790, 785, 760, 760, 770

a. Construct the ANOVA table for these data.

b. Compute the p-value for the null hypothesis that all three sizes of vehicles produce the same level of noise on average.

11. Assume that the null hypothesis H_0 in (11.6.5) is false. Prove that the expected value of S^2_{Betw} is $(p-1)\sigma^2 + \sum_{i=1}^{p} n_i(\mu_i - \overline{\mu})^2$, where $\overline{\mu} = \frac{1}{n}\sum_{i=1}^{p} n_i\mu_i$.

12. Prove that the level α_0 likelihood ratio test of hypotheses (11.6.5) in the one-way layout is to reject H_0 if $U^2 > F^{-1}_{p-1,\,n-p}(1-\alpha_0)$. *Hint:* First, partition $\sum_{j=1}^{n_i}(y_{ij} - \mu_i)^2$ in a manner similar to Eq. (11.6.8). Then, replace $\overline{Y}_{++}$ by a constant, say, μ, in the formula for S^2_{Tot}, and partition the result in a manner similar to Eq. (11.6.7). There will be one extra term.

13. Suppose that the null hypothesis in (11.6.5) is true. Prove that $S^2_{\text{Tot}}/\sigma^2$ has the χ^2 distribution with $n-1$ degrees of freedom.

14. A popular alternative parameterization of the one-way layout is the following. Let $\mu = \frac{1}{n}\sum_{i=1}^{p} n_i\mu_i$, and define $\alpha_i = \mu_i - \mu$. This makes $E(Y_{ij}) = \mu + \alpha_i$.

a. Prove that $\sum_{i=1}^{p} \alpha_i = 0$.

b. Prove that the M.L.E. of α_i is $\overline{Y}_{i+} - \overline{Y}_{++}$.

c. Prove that the null hypothesis H_0 in (11.6.5) is equivalent to $\alpha_1 = \cdots = \alpha_p = 0$.

d. Prove that the mean of S^2_{Betw} is $(p-1)\sigma^2 + \sum_{i=1}^{p} n_i\alpha_i^2$.

⋆ 11.7 The Two-Way Layout

In Sec. 11.6, we learned how to analyze several samples that differed in some characteristic. For example, we analyzed data collected from hot dogs that differed by the type of meat from which they were made. Suppose that, in addition to differing by the type of meat, the hot dogs had also differed by being labeled either "low fat" or not. This would have given us two different characteristics to form the basis for comparisons. In this section, we shall study how to analyze data consisting of observations that differ on two characteristics.

The Two-Way Layout with One Observation in Each Cell

Example 11.7.1

Radioactive Isotope in Milk. Suppose that in an experiment to measure the concentration of a certain radioactive isotope in milk, specimens of milk are obtained from four different dairies, and the concentration of the isotope in each specimen is measured by three different methods. If we let Y_{ij} denote the measurement that is made for the specimen from the ith dairy by using the jth method, for $i = 1, 2, 3, 4$ and $j = 1, 2, 3$, then in this example there will be a total of 12 measurements. There are two main questions of interest in this example. The first is whether the concentration of the isotope is the same in the milk of all four dairies. The second question is whether the three different methods produce concentration measurements that appear to differ. ◀

A problem of the type in Example 11.7.1, in which the value of the random variable being observed is affected by two factors, is called a *two-way layout*. In the general two-way layout, there are two factors, which we shall call A and B. We shall assume that there are I possible different values, or different *levels*, of factor A, and that there are J possible different values, or different *levels*, of factor B. For $i = 1, \ldots, I$ and $j = 1, \ldots, J$, an observation Y_{ij} of the variable being studied is obtained when factor A has the value i and factor B has the value j. If the IJ observations are arranged in a matrix as in Table 11.20, then Y_{ij} is the observation in the (i, j) cell of the matrix.

We shall continue to make the assumptions of the general linear model for the two-way layout. Thus, we shall assume that all the observations Y_{ij} are independent, each observation has a normal distribution, and all the observations have the same variance σ^2. In this section, we specialize the assumption about the mean $E(Y_{ij})$ as follows: We shall assume not only that $E(Y_{ij})$ depends on the values i and j of the two factors, but also that there exist numbers $\theta_1, \ldots, \theta_I$ and $\psi_1, \ldots, \psi_J$ such that

$$E(Y_{ij}) = \theta_i + \psi_j \quad \text{for } i = 1, \ldots, I \quad \text{and } j = 1, \ldots, J. \tag{11.7.1}$$

Thus, Eq. (11.7.1) states that the value of $E(Y_{ij})$ is the sum of the following two effects: an effect θ_i due to factor A having the value i, and an effect ψ_j due to factor B

Table 11.20 Generic data for two-way layout

	Factor B			
Factor A	1	2	$\cdots$	J
1	Y_{11}	Y_{12}	$\cdots$	Y_{1J}
2	Y_{21}	Y_{22}		Y_{2J}
$\vdots$				
I	Y_{I1}	Y_{I2}		Y_{IJ}

having the value j. For this reason, the assumption that $E(Y_{ij})$ has the form given in Eq. (11.7.1) is called an *assumption of additivity* of the effects of the factors.

The meaning of the assumption of additivity can be clarified by the following example. Consider the sale of I different magazines at J different newsstands. Suppose that a particular newsstand sells on the average 30 more copies per week of magazine 1 than of magazine 2. Then by the assumption of additivity, it must also be true that each of the other $J-1$ newsstands sells on the average 30 more copies per week of magazine 1 than of magazine 2. Similarly, suppose that the sales of a particular magazine are on the average 50 more copies per week at newsstand 1 than at newsstand 2. Then by the assumption of additivity, it must also be true that the sales of each of the other $I-1$ magazines are on the average 50 more copies per week at newsstand 1 than at newsstand 2. The assumption of additivity is a very restrictive assumption because it does not allow for the possibility that a particular magazine may sell unusually well at some particular newsstand. In Sec. 11.8, we shall consider models in which we do not make the assumption of additivity.

Even though we assume in the general two-way layout that the effects of the factors A and B are additive, the numbers θ_i and ψ_j that satisfy Eq. (11.7.1) are not uniquely defined. We can add an arbitrary constant c to each of the numbers $\theta_1, \ldots, \theta_I$ and subtract the same constant c from each of the numbers $\psi_1, \ldots, \psi_J$ without changing the value of $E(Y_{ij})$ for any of the IJ observations. Hence, it does not make sense to try to estimate the value of θ_i or ψ_j from the given observations, since neither θ_i nor ψ_j is uniquely defined. In order to avoid this difficulty, we shall express $E(Y_{ij})$ in terms of different parameters. The following assumption is equivalent to the assumption of additivity.

We shall assume that there exist numbers $\mu, \alpha_1, \ldots, \alpha_I$, and $\beta_1, \ldots, \beta_J$ such that

$$\sum_{i=1}^{I} \alpha_i = 0 \quad \text{and} \quad \sum_{j=1}^{J} \beta_j = 0, \tag{11.7.2}$$

and

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j \quad \text{for } i = 1, \ldots, I \text{ and } j = 1, \ldots, J. \tag{11.7.3}$$

There is an advantage in expressing $E(Y_{ij})$ in this way. If the values of $E(Y_{ij})$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ are a set of numbers that satisfy Eq. (11.7.1) for *some* set of values of $\theta_1, \ldots, \theta_I$ and $\psi_1, \ldots, \psi_J$, then there exists a *unique* set of values of $\mu, \alpha_1, \ldots, \alpha_I$, and $\beta_1, \ldots, \beta_J$ that satisfy Eqs. (11.7.2) and (11.7.3) (see Exercise 3).

The parameter μ is called the *overall mean*, or the *grand mean*, since it follows from Eqs. (11.7.2) and (11.7.3) that

$$\mu = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} E(Y_{ij}). \tag{11.7.4}$$

The parameters $\alpha_1, \ldots, \alpha_I$ are called the *effects of factor A*, and the parameters $\beta_1, \ldots, \beta_J$ are called the *effects of factor B*.

It follows from Eq. (11.7.2) that $\alpha_I = -\sum_{i=1}^{I-1}\alpha_i$ and $\beta_J = -\sum_{j=1}^{J-1}\beta_j$. Hence, each expectation $E(Y_{ij})$ in Eq. (11.7.3) can be expressed as a particular linear combination of the $I + J - 1$ parameters $\mu, \alpha_1, \ldots, \alpha_{I-1}$, and $\beta_1, \ldots, \beta_{J-1}$. Therefore, if we regard the IJ observations as elements of a single long IJ-dimensional vector, then the two-way layout satisfies the conditions of the general linear model. In a practical problem, however, it is not convenient to actually replace α_I and β_J with their expressions in terms of the other α_i's and β_j's, because this replacement would destroy the symmetry that is present in the experiment among the different levels of each factor.

Estimating the Parameters

The following result is straightforward, but tedious, to prove.

Theorem 11.7.1 Define

$$\overline{Y}_{i+} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij} \quad \text{for } i = 1, \ldots, I,$$
$$\overline{Y}_{+j} = \frac{1}{I}\sum_{i=1}^{I} Y_{ij} \quad \text{for } j = 1, \ldots, J, \tag{11.7.5}$$
$$\overline{Y}_{++} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij} = \frac{1}{I}\sum_{i=1}^{I}\overline{Y}_{i+} = \frac{1}{J}\sum_{j=1}^{J}\overline{Y}_{+j}.$$

Then the M.L.E.'s (and least-squares estimators) of μ, $\alpha_1, \ldots, \alpha_I$, and $\beta_1, \ldots, \beta_J$ are as follows:

$$\hat{\mu} = \overline{Y}_{++},$$
$$\hat{\alpha}_i = \overline{Y}_{i+} - \overline{Y}_{++} \quad \text{for } i = 1, \ldots, I, \tag{11.7.6}$$
$$\hat{\beta}_j = \overline{Y}_{+j} - \overline{Y}_{++} \quad \text{for } j = 1, \ldots, J.$$

The M.L.E. of σ^2 will be

$$\hat{\sigma}^2 = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2 = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \hat{Y}_{ij})^2. \quad \blacksquare$$

It is easily verified (see Exercise 6) that $\sum_{i=1}^{I}\hat{\alpha}_i = \sum_{j=1}^{J}\hat{\beta}_j = 0$; $E(\hat{\mu}) = \mu$; $E(\hat{\alpha}_i) = \alpha_i$ for $i = 1, \ldots, I$; and $E(\hat{\beta}_j) = \beta_j$ for $j = 1, \ldots, J$. Because $E(Y_{ij}) = \mu + \alpha_i + \beta_j$, the M.L.E. of $E(Y_{ij})$ is

$$\hat{Y}_{ij} = \overline{Y}_{i+} + \overline{Y}_{+j} - \overline{Y}_{++} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j,$$

which is also called the *fitted value* for Y_{ij}.

Example 11.7.2 Radioactive Isotope in Milk. Consider again Example 11.7.1. Suppose that the concentrations of the radioactive isotope measured in picocuries per liter by three different methods in specimens of milk from four dairies are as shown in Table 11.21. From

Table 11.21 Data for Example 11.7.2

	Method		
Dairy	1	2	3
1	6.4	3.2	6.9
2	8.5	7.8	10.1
3	9.3	6.0	9.6
4	8.8	5.6	8.4

Table 11.22 Fitted values for observations in Example 11.7.2

	Method		
Dairy	1	2	3
1	6.2	3.6	6.7
2	9.5	6.9	10.0
3	9.0	6.4	9.5
4	8.3	5.7	8.8

Table 11.21, the row averages are $\overline{Y}_{1+} = 5.5$, $\overline{Y}_{2+} = 8.8$, $\overline{Y}_{3+} = 8.3$, and $\overline{Y}_{4+} = 7.6$; the column averages are $\overline{Y}_{+1} = 8.25$, $\overline{Y}_{+2} = 5.65$, and $\overline{Y}_{+3} = 8.75$; and the average of all the observations is $\overline{Y}_{++} = 7.55$. Hence, by Eq. (11.7.6), the values of the M.L.E.'s are $\hat{\mu} = 7.55$, $\hat{\alpha}_1 = -2.05$, $\hat{\alpha}_2 = 1.25$, $\hat{\alpha}_3 = 0.75$, $\hat{\alpha}_4 = 0.05$, $\hat{\beta}_1 = 0.70$, $\hat{\beta}_2 = -1.90$, and $\hat{\beta}_3 = 1.20$.

The fitted values $\hat{Y}_{ij}$ for all of the observations are given in Table 11.22. By comparing the observed values in Table 11.21 with the fitted values in Table 11.22, we see that the differences between corresponding terms are generally small. These small differences indicate that the model used in the two-way layout, which assumes the additivity of the effects of the two factors, provides a good fit for the observed values. Finally, it is found from Tables 11.21 and 11.22 that

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \hat{Y}_{ij})^2 = 2.74.$$

Hence, by Theorem 11.7.1, $\hat{\sigma}^2 = 2.74/12 = 0.228$. ◀

Partitioning the Sum of Squares

We shall partition the total sum of squares in much the same way that we did in Sec. 11.6. Begin with

$$S^2_{\text{Tot}} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_{++})^2. \tag{11.7.7}$$

We shall now partition the sum of squares S^2_{Tot} into three smaller sums of squares. Each of these smaller sums of squares will be associated with a certain type of variation among the observations Y_{ij}. Each of them (divided by σ^2) will have a χ^2 distribution if certain null hypotheses are true, and they will be mutually independent whether or not the null hypotheses are true. Therefore, just as in the one-way layout, we can construct tests of certain null hypotheses based on an analysis of variance, that is, on an analysis of these different types of variation.

Theorem 11.7.2 Partitioning the Sum of Squares. Let S^2_{Tot} be as defined in Eq. (11.7.7). Then

$$S^2_{\text{Tot}} = S^2_{\text{Resid}} + S^2_A + S^2_B, \tag{11.7.8}$$

where

$$S^2_{\text{Resid}} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_{i+} - \overline{Y}_{+j} + \overline{Y}_{++})^2,$$

$$S^2_A = J\sum_{i=1}^{I}(\overline{Y}_{i+} - \overline{Y}_{++})^2,$$

$$S^2_B = I\sum_{j=1}^{J}(\overline{Y}_{+j} - \overline{Y}_{++})^2.$$

Furthermore, $S^2_{\text{Resid}}/\sigma^2$ has the χ^2 distribution with $(I-1)(J-1)$ degrees of freedom, and the three component sums of squares are mutually independent.

Proof We shall begin by rewriting S^2_{Tot} as follows:

$$S^2_{\text{Tot}} = \sum_{i=1}^{I}\sum_{j=1}^{J}[(Y_{ij} - \overline{Y}_{i+} - \overline{Y}_{+j} + \overline{Y}_{++}) + (\overline{Y}_{i+} - \overline{Y}_{++}) + (\overline{Y}_{+j} - \overline{Y}_{++})]^2. \tag{11.7.9}$$

By expanding the right side of Eq. (11.7.9), we obtain (see Exercise 8) Eq. (11.7.8).

It can be shown that the random variables S^2_{Resid}, S^2_{A}, and S^2_{B} are independent. (See Exercise 9 for a related result.) Furthermore, it can be shown that S^2_{Resid} has the χ^2 distribution with $IJ - (I + J - 1) = (I-1)(J-1)$ degrees of freedom. ■

It is easy to see that S^2_A measures the variation of the sample means for the different levels of factor A around the overall sample mean. Similarly, S^2_B measures the variation of the sample means for the different levels of factor B around the overall sample mean. By using relations (11.7.6), we can rewrite S^2_{Resid} as

$$S^2_{\text{Resid}} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \hat{Y}_{ij})^2.$$

This makes it clear that S^2_{Resid} measures the residual variation, that is, the variation between the observations not explained by the model. The partitioning is summarized in Table 11.23, which is the ANOVA table for the two-way layout. As in the case of the one-way layout, the degrees of freedom will turn out to be degrees of freedom for various χ^2 random variables when certain null hypotheses are true.

Table 11.23 General ANOVA table for two-way layout

Source of variation	Degrees of freedom	Sum of squares	Mean square
Factor A	$I-1$	S_A^2	$S_A^2/(I-1)$
Factor B	$J-1$	S_B^2	$S_B^2/(J-1)$
Residuals	$(I-1)(J-1)$	S_{Resid}^2	$S_{\text{Resid}}^2/[(I-1)(J-1)]$
Total	$IJ-1$	S_{Tot}^2	

Table 11.24 ANOVA table Example 11.7.3

Source of variation	Degrees of freedom	Sum of squares	Mean square
Dairy	3	18.99	6.33
Method	2	22.16	11.08
Residuals	6	2.74	0.4567
Total	11	43.89	

Example 11.7.3 Radioactive Isotope in Milk. Using the estimates calculated in Example 11.7.2, we can compute the ANOVA table in Table 11.24. After we develop appropriate test statistics, we can use Table 11.24 to test hypotheses about the effects of the two factors. ◀

Testing Hypotheses

Example 11.7.4 Radioactive Isotope in Milk. Consider again the situation described in Example 11.7.2 involving four dairies and three measurement methods. We might be interested in testing that, for each of the three methods of measurement, the distributions of concentration of isotope do not differ from one dairy to the next. If we regard the dairy as factor A and the measurement method as factor B, then the hypothesis that $\alpha_i = 0$ for $i = 1, \dots, I$ means that for each method of measurement, the concentration of the isotope has the same distribution for all four dairies. In other words, there are no differences among the dairies. Alternatively, we might be interested in testing the hypothesis that, for each dairy, the three methods of measurement all produce the same distribution of concentration of isotope. For this case, the hypothesis that $\beta_j = 0$ for $j = 1, \dots, J$ means that for each dairy, the three methods of measurement yield the same distribution for the concentration of the isotope. However, this hypothesis does not state that regardless of which of the three different methods is applied to a particular specimen of milk, the same value would be obtained. Because of the inherent variability of the measurements, the hypothesis states only that the values yielded by the three methods have the same normal distribution. ◀

In a problem involving a two-way layout, we are often interested in testing the hypothesis that one or both of the factors has no effect on the distribution of the observations. In other words, we are often interested either in testing the hypothesis that all of the effects $\alpha_1, \dots, \alpha_I$ of factor A are equal to 0 or in testing the hypothesis that all of the effects $\beta_1, \dots, \beta_J$ of factor B are equal to 0 or in testing that all of the α_i and β_j are 0. For the remainder of the discussion of testing hypotheses, it will be useful to define

$$\sigma' = \left(\frac{S^2_{\text{Resid}}}{(I-1)(J-1)} \right)^{1/2}. \tag{11.7.10}$$

Theorem 11.7.3 Consider the following hypotheses:

$$\begin{aligned} H_0&: \quad \alpha_i = 0 \quad \text{for } i = 1, \dots, I, \\ H_1&: \quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{11.7.11}$$

If H_0 is true, then the following random variable has the F distribution with $I-1$ and $(I-1)(J-1)$ degrees of freedom:

$$U_A^2 = \frac{S_A^2}{(I-1)\sigma'^2}. \tag{11.7.12}$$

Similarly, suppose next that the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad \beta_j = 0 \quad \text{for } j = 1, \dots, J, \\ H_1&: \quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{11.7.13}$$

When the null hypothesis H_0 is true, the following statistic has the F distribution with $J-1$ and $(I-1)(J-1)$ degrees of freedom:

$$U_B^2 = \frac{S_B^2}{(J-1)\sigma'^2}. \tag{11.7.14}$$

Finally, suppose that the following hypotheses are to be tested:

$$\begin{aligned} H_0&: \quad \alpha_i = 0 \text{ for } i = 1, \dots, I, \quad \text{and} \quad \beta_j = 0 \text{ for } j = 1, \dots, J, \\ H_1&: \quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{11.7.15}$$

When the null hypothesis H_0 is true, the following statistic has the F distribution with $I+J-2$ and $(I-1)(J-1)$ degrees of freedom:

$$U_{A+B}^2 = \frac{S_A^2 + S_B^2}{(I+J-2)\sigma'^2}. \tag{11.7.16}$$

For each case above, a level α_0 test of the hypotheses is to reject H_0 if the corresponding statistic (U_A^2, U_B^2, or U_{A+B}^2) is at least as large as the $1-\alpha_0$ quantile of the correpsonding F distribution.

Proof We shall prove the claim for hypotheses (11.7.11). The proof for hypotheses (11.7.13) is virtually identical. The proof for hypotheses (11.7.15) is similar and is left for Exercise 16. Since $\sum_{j=1}^{J} \beta_j = 0$, we conclude that $\overline{Y}_{i+}$ has the normal distribution with mean μ and variance σ^2/J for each $i = 1, \dots, I$. Since the $\overline{Y}_{i+}$ are independent and $\overline{Y}_{++}$ is the average of $\overline{Y}_{1+}, \dots, \overline{Y}_{I+}$, Theorem 8.3.1 says that the following

random variable has the χ^2 distribution with $I-1$ degrees of freedom:

$$\frac{J}{\sigma^2}\sum_{i=1}^{I}(\overline{Y}_{i+}-\overline{Y}_{++})^2=\frac{S_A^2}{\sigma^2}.$$

Since $S^2_{\text{Resid}}/\sigma^2$ has the χ^2 distribution with $(I-1)(J-1)$ degrees of freedom, we now conclude that

$$\frac{S_A^2/(I-1)}{S^2_{\text{Resid}}/[(I-1)(J-1)]} \tag{11.7.17}$$

has the F distribution with $I-1$ and $(I-1)(J-1)$ degrees of freedom. It is easy to see that the random variable in (11.7.17) is the same as U_A^2 defined in Eq. (11.7.12). Let $F^{-1}_{I-1,(I-1)(J-1)}(1-\alpha_0)$ denote the $1-\alpha_0$ quantile of the F distribution with $I-1$ and $(I-1)(J-1)$ degrees of freedom. Let δ be the test that rejects H_0 if $U_A^2\geq F^{-1}_{I-1,(I-1)(J-1)}(1-\alpha_0)$, and let $\pi(\theta|\delta)$ be its power function for each parameter vector θ. Since U_A^2 has the stated F distribution for all parameter vectors θ that are consistent with H_0, it follows that for each such θ, $\pi(\theta|\delta)=\alpha_0$, and δ is a level α_0 test. ■

Notice that U_A^2 in Theorem 11.7.3 is the ratio of the factor A mean square to the residuals mean square in Table 11.23. When the null hypothesis H_0 in (11.7.12) is not true, the value of $\alpha_i=E(\overline{Y}_{i+}-\overline{Y}_{++})$ is not 0 for at least one value of i. Hence, the expectation of the numerator of U_A^2 will be larger than it would be when H_0 is true. (See Exercise 1.) The distribution of the denominator of U_A^2 remains the same regardless of whether H_0 is true. It can also be shown that the test in Theorem 11.7.3 is also the level α_0 likelihood ratio test procedure for the hypotheses (11.7.11).

Example 11.7.5 Testing for Differences among the Dairies. Suppose now that it is desired to use the observed values in Table 11.21 to test the hypothesis that there are no differences among the dairies, that is, to test the hypotheses (11.7.11). In this example, the statistic U_A^2 defined by Eq. (11.7.12) has the F distribution with three and six degrees of freedom. Using the ANOVA table in Table 11.24, we find that $U_A^2=6.33/0.4567=13.86$. The corresponding p-value is smaller than 0.025, the smallest value in the tables in this book. Using statistical software, we compute the p-value to be about 0.004. So the hypothesis that there are no differences among the dairies would be rejected at all levels of significance of 0.004 or more. ◄

Example 11.7.6 Testing for Differences among the Methods of Measurement. Suppose next that it is desired to use the observed values in Table 11.21 to test the hypothesis that each of the effects of the different methods of measurement is equal to 0, that is, to test the hypotheses (11.7.13). In this example, the statistic U_B^2 defined by Eq. (11.7.14) has the F distribution with two and six degrees of freedom. Using the ANOVA table in Table 11.24, we find that $U_B^2=11.08/0.4567=24.26$. The p-value corresponding to this observation is about 0.001, so the hypothesis that there are no differences among the methods would be rejected at all levels of significance greater than 0.001. ◄

Summary

The two-way layout can be considered as a general linear model, but the hypotheses of interest concern more than one linear combination of the regression coefficients. An ANOVA table was developed for the two-way layout that can be used for forming test statistics for various hypotheses. When we have only one observation at each

combination of factor levels, we assume that the effects of the two factors are additive. Then we can test the two null hypotheses that each of the two factors make no difference to the means of the observations. These tests make use of the test statistics U_A^2 in Eq. (11.7.12) and U_B^2 in Eq. (11.7.14). If the corresponding null hypotheses are true, each of these statistics has an F distribution.

Exercises

1. Suppose that the null hypothesis H_0 in (11.7.11) is false. Show that $E(S_A^2) = (I-1)\sigma^2 + J\sum_{i=1}^I \alpha_i^2$.

2. Consider a two-way layout in which the values of $E(Y_{ij})$ for $i = 1, \dots, I$ and $j = 1, \dots, J$ are as given in each of the following four matrices. For each matrix, state whether the effects of the factors are additive.

a.

	Factor B	
Factor A	1	2
1	5	7
2	10	14

b.

	Factor B	
Factor A	1	2
1	3	6
2	4	7

c.

	Factor B			
Factor A	1	2	3	4
1	3	−1	0	3
2	8	4	5	8
3	4	0	1	4

d.

	Factor B			
Factor A	1	2	3	4
1	1	2	3	4
2	2	4	6	8
3	3	6	9	12

3. Show that if the effects of the factors in a two-way layout are additive, then there exist unique numbers μ, $\alpha_1, \dots, \alpha_I$, and $\beta_1, \dots, \beta_J$ that satisfy Eqs. (11.7.2) and (11.7.3). *Hint:* Let μ be the average of all $\theta_i + \psi_j$ values, let α_i equal θ_i minus the average of the θ_i's, and similarly for β_j.

4. Suppose that in a two-way layout, with $I = 2$ and $J = 2$, the values of $E(Y_{ij})$ are as given in part (b) of Exercise 2. Determine the values of $\mu, \alpha_1, \alpha_2, \beta_1$, and β_2 that satisfy Eqs. (11.7.2) and (11.7.3).

5. Suppose that in a two-way layout, with $I = 3$ and $J = 4$, the values of $E(Y_{ij})$ are as given in part (c) of Exercise 2. Determine the values of $\mu, \alpha_1, \alpha_2, \alpha_3$, and $\beta_1, \dots, \beta_4$ that satisfy Eqs. (11.7.2) and (11.7.3).

6. Verify that if $\hat{\mu}$, $\hat{\alpha}_i$, and $\hat{\beta}_j$ are defined by Eq. (11.7.6), then $\sum_{i=1}^I \hat{\alpha}_i = \sum_{j=1}^J \hat{\beta}_j = 0$; $E(\hat{\mu}) = \mu$; $E(\hat{\alpha}_i) = \alpha_i$ for $i = 1, \dots, I$; and $E(\hat{\beta}_j) = \beta_j$ for $j = 1, \dots, J$.

7. Show that if $\hat{\mu}$, $\hat{\alpha}_i$, and $\hat{\beta}_j$ are defined by Eq. (11.7.6), then

$$\begin{aligned}
\text{Var}(\hat{\mu}) &= \frac{1}{IJ}\sigma^2, \\
\text{Var}(\hat{\alpha}_i) &= \frac{I-1}{IJ}\sigma^2 \quad \text{for } i = 1, \dots, I, \\
\text{Var}(\hat{\beta}_j) &= \frac{J-1}{IJ}\sigma^2 \quad \text{for } j = 1, \dots, J.
\end{aligned}$$

8. Show that the right sides of Eqs. (11.7.9) and (11.7.8) are equal.

9. Show that in a two-way layout, for all values of i, j, i', and j' (i and $i' = 1, \dots, I$; j and $j' = 1, \dots, J$), the following four random variables W_1, W_2, W_3, and W_4 are uncorrelated with one another:

$$\begin{aligned}
&W_1 = Y_{ij} - \overline{Y}_{i+} - \overline{Y}_{+j} + \overline{Y}_{++}, \\
&W_2 = \overline{Y}_{i'+} - \overline{Y}_{++}, \quad W_3 = \overline{Y}_{+j'} - \overline{Y}_{++}, \\
&W_4 = \overline{Y}_{++}.
\end{aligned}$$

10. Show that

$$\sum_{i=1}^I (\overline{Y}_{i+} - \overline{Y}_{++})^2 = \sum_{i=1}^I \overline{Y}_{i+}^2 - I\overline{Y}_{++}^2$$

and

$$\sum_{j=1}^J (\overline{Y}_{+j} - \overline{Y}_{++})^2 = \sum_{j=1}^J \overline{Y}_{+j}^2 - J\overline{Y}_{++}^2.$$

11. Show that

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \overline{Y}_{i+} - \overline{Y}_{+j} + \overline{Y}_{++})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}^2 - J\sum_{i=1}^{I}\overline{Y}_{i+}^2 - I\sum_{j=1}^{J}\overline{Y}_{+j}^2 + IJ\overline{Y}_{++}^2.$$

12. In a study to compare the reflective properties of various paints and various plastic surfaces, three different types of paint were applied to specimens of five different types of plastic surfaces. Suppose that the observed results in appropriate coded units were as shown in Table 11.25. Determine the values of $\hat{\mu}$, $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$, and $\hat{\beta}_1, \dots, \hat{\beta}_5$.

Table 11.25 Data for Exercises 12–15

	Type of surface				
Type of paint	1	2	3	4	5
1	14.5	13.6	16.3	23.2	19.4
2	14.6	16.2	14.8	16.8	17.3
3	16.2	14.0	15.5	18.7	21.0

13. For the conditions of Exercise 12 and the data in Table 11.25, determine the value of the least-squares estimate of $E(Y_{ij})$ for $i = 1, 2, 3$, and $j = 1, \dots, 5$, and determine the value of $\hat{\sigma}^2$.

14. For the conditions of Exercise 12 and the data in Table 11.25, test the hypothesis that the reflective properties of the three different types of paint are the same.

15. For the conditions of Exercise 12 and the data in Table 11.25, test the hypothesis that the reflective properties of the five different types of plastic surfaces are the same.

16. Prove the claim in Theorem 11.7.3 about the distribution of U^2_{A+B}.

⋆ 11.8 The Two-Way Layout with Replications

Suppose that we obtain more than one observation in each cell of a two-way layout. In addition to testing hypotheses about the separate effects of the two factors, we can also test the hypothesis that the additivity assumption (11.7.3) holds. However, the interpretations of the separate effects of the two factors are more complicated if the additivity assumption fails. When the additivity assumption fails, we say that there is interaction between the two factors.

The Two-Way Layout with K Observations in Each Cell

Example 11.8.1

Gasoline Consumption. Suppose that an experiment is carried out by an automobile manufacturer to investigate whether a certain device, installed on an automobile, affects the amount of gasoline consumed by the automobile. The manufacturer produces three different models of automobiles, namely, a compact model, an intermediate model, and a standard model. Five cars of each model, which were equipped with this device, were driven over a fixed route through city traffic, and the gasoline consumption of each car was measured. Also, five cars of each model, which were not equipped with this device, were driven over the same route, and the gasoline consumption of each of these cars was measured. The results, in liters of gasoline consumed, are given in Table 11.26.

The same sorts of questions that arose in Sec. 11.7 arise here. For example, are the mean gasoline consumptions different for cars with and without the device? Are the mean gasoline consumptions different for the three car models? An additional question can be addressed in an example like this in which there are multiple obser-

Table 11.26 Data for Example 11.8.1

	Compact model	Intermediate model	Standard model
Equipped with device	8.3	9.2	11.6
	8.9	10.2	10.2
	7.8	9.5	10.7
	8.5	11.3	11.9
	9.4	10.4	11.0
Not equipped with device	8.7	8.2	12.4
	10.0	10.6	11.7
	9.7	10.1	10.0
	7.9	11.3	11.1
	8.4	10.8	11.8

vations under each combination of factors. We can ask whether the effect (if any) of the device is different for the different car models. ◀

We shall continue to consider problems of ANOVA involving a two-way layout. Now, however, instead of having just a single observation Y_{ij} for each combination of i and j, we shall have K independent observations Y_{ijk} for $k = 1, \dots, K$. In other words, instead of having just one observation in each cell of Table 11.20, we have K i.i.d. observations. The K observations in each cell are obtained under similar experimental conditions and are called *replications*. The total number of observations in this two-way layout with replications is IJK. We continue to assume that all the observations are independent, each observation has a normal distribution, and all the observations have the same variance σ^2.

We shall let θ_{ij} denote the mean of each of the K observations in the (i, j) cell. Thus, for $i = 1, \dots, I$; $j = 1, \dots, J$; and $k = 1, \dots, K$, we have

$$E(Y_{ijk}) = \theta_{ij}. \tag{11.8.1}$$

In a two-way layout with replications, we shall no longer assume, as we did in Sec. 11.7, that the effects of the two factors are additive. Here we can assume that the expectations θ_{ij} are arbitrary numbers. As we shall see later in this section, we can then test the hypothesis that the effects are additive.

It is easy to verify that the M.L.E., or least-squares estimator, of θ_{ij} is simply the sample mean of the K observations in the (i, j) cell. Thus,

$$\hat{\theta}_{ij} = \frac{1}{K}\sum_{k=1}^{K} Y_{ijk} = \overline{Y}_{ij+}. \tag{11.8.2}$$

The M.L.E. of σ^2 is therefore

$$\hat{\sigma}^2 = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \overline{Y}_{ij+})^2. \tag{11.8.3}$$

In order to identify and discuss the effects of the two factors, and to examine the possibility that these effects are additive, it is helpful to replace the parameters θ_{ij}, for $i = 1, \dots, I$ and $j = 1, \dots, J$, with a new set of parameters μ, α_i, β_j, and γ_{ij}. These new parameters are defined by the following relations:

$$\theta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad \text{for } i = 1, \dots, I \text{ and } j = 1, \dots, J, \tag{11.8.4}$$

and

$$\begin{aligned} &\sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0, \\ &\sum_{i=1}^{I} \gamma_{ij} = 0 \quad \text{for } j = 1, \dots, J, \\ &\sum_{j=1}^{J} \gamma_{ij} = 0 \quad \text{for } i = 1, \dots, I. \end{aligned} \tag{11.8.5}$$

It can be shown (see Exercise 1) that corresponding to each set of numbers θ_{ij} for $i = 1, \dots, I$ and $j = 1, \dots, J$, there exist unique numbers μ, α_i, β_j, and γ_{ij} that satisfy Eqs. (11.8.4) and (11.8.5).

The parameter μ is called the *overall mean* or the *grand mean*. The parameters $\alpha_1, \dots, \alpha_I$ are called the *main effects of factor A*, and the parameters $\beta_1, \dots, \beta_J$ are called the *main effects of factor B*. The parameters γ_{ij}, for $i = 1, \dots, I$ and $j = 1, \dots, J$, are called the *interactions*. It can be seen from Eqs. (11.8.1) and (11.8.4) that the effects of the factors A and B are additive if and only if all the interactions vanish, that is, if and only if $\gamma_{ij} = 0$ for every combination of values of i and j.

The notation that has been developed in Sections 11.6 and 11.7 will again be used here. We shall replace a subscript of Y_{ijk} with a plus sign to indicate that we have summed the values of Y_{ijk} over all possible values of that subscript. If we have made two or three summations, we shall use two or three plus signs. We shall then place a bar over Y to indicate that we have divided this sum by the number of terms in the summation and have thereby obtained an average of the values of Y_{ijk} for the subscript or subscripts involved in the summation. For example,

$$\overline{Y}_{ij+} = \frac{1}{K} \sum_{k=1}^{K} Y_{ijk},$$

$$\overline{Y}_{+j+} = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} Y_{ijk},$$

and $\overline{Y}_{+++}$ denotes the average of all IJK observations.

Similar logic to that used in the proof of Theorem 11.2.1 can be used to prove the following result. The details are left to Exercises 2 and 5).

Theorem 11.8.1 The M.L.E.'s (and least-squares estimators) of μ, α_i, and β_j are as follows:

$$\begin{aligned} &\hat{\mu} = \overline{Y}_{+++}, \\ &\hat{\alpha}_i = \overline{Y}_{i++} - \overline{Y}_{+++} \quad \text{for } i = 1, \dots, I, \\ &\hat{\beta}_j = \overline{Y}_{+j+} - \overline{Y}_{+++} \quad \text{for } j = 1, \dots, J. \end{aligned} \tag{11.8.6}$$

Also, for $i = 1, \dots, I$ and $j = 1, \dots, J$,

Table 11.27 Cell averages in Example 11.8.2

	Compact model	Intermediate model	Standard model	Average for row
Equipped with device	$\overline{Y}_{11+} = 8.58$	$\overline{Y}_{12+} = 10.12$	$\overline{Y}_{13+} = 11.08$	$\overline{Y}_{1++} = 9.9267$
Not equipped with device	$\overline{Y}_{21+} = 8.94$	$\overline{Y}_{22+} = 10.20$	$\overline{Y}_{23+} = 11.40$	$\overline{Y}_{2++} = 10.1800$
Average for column	$\overline{Y}_{+1+} = 8.76$	$\overline{Y}_{+2+} = 10.16$	$\overline{Y}_{+3+} = 11.24$	$\overline{Y}_{+++} = 10.0533$

$$\begin{aligned}\hat{\gamma}_{ij} &= \overline{Y}_{ij+} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j) \\ &= \overline{Y}_{ij+} - \overline{Y}_{i++} - \overline{Y}_{+j+} + \overline{Y}_{+++}.\end{aligned} \tag{11.8.7}$$

Also, for all values of i and j, $E(\hat{\mu}) = \mu$, $E(\hat{\alpha}_i) = \alpha_i$, $E(\hat{\beta}_j) = \beta_j$, and $E(\hat{\gamma}_{ij}) = \gamma_{ij}$.

Example 11.8.2

Gasoline Consumption. In Example 11.8.1, let the A factor be the device, and let the B factor be the car model. Then we have $I = 2$, $J = 3$, and $K = 5$. The average value $\overline{Y}_{ij+}$ for each of the six cells in Table 11.26 is presented in Table 11.27, which also gives the average value $\overline{Y}_{i++}$ for each of the two rows, the average value $\overline{Y}_{+j+}$ for each of the three columns, and the average value $\overline{Y}_{+++}$ of all 30 observations.

It follows from Table 11.27 and Eqs. (11.8.6) and (11.8.7) that the values of the M.L.E.'s, or least-squares estimators, in this example are

$$\begin{array}{llllll}\hat{\mu} = 10.0533, & \hat{\alpha}_1 = -0.1267, & \hat{\alpha}_2 = 0.1267, \\ \hat{\beta}_1 = -1.2933, & \hat{\beta}_2 = 0.1067, & \hat{\beta}_3 = 1.1867, \\ \hat{\gamma}_{11} = -0.0533, & \hat{\gamma}_{12} = 0.0867, & \hat{\gamma}_{13} = -0.0333, \\ \hat{\gamma}_{21} = 0.0533, & \hat{\gamma}_{22} = -0.0867, & \hat{\gamma}_{23} = 0.0333.\end{array}$$

In this example, the estimates of the interactions $\hat{\gamma}_{ij}$ are small for all values of i and j. ◀

Partitioning the Sum of Squares

Consider now the total sum of squares,

$$S^2_{\text{Tot}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (Y_{ijk} - \overline{Y}_{+++})^2. \tag{11.8.8}$$

We shall now indicate how S^2_{Tot} can be partitioned into four smaller independent sums of squares, each of which is associated with a particular type of variation among the observations. Under various null hypotheses, each sum of squares (divided by σ^2) will have a χ^2 distribution.

Theorem 11.8.2

Let S^2_{Tot} be as defined in Eq. (11.8.8). Then

$$S^2_{\text{Tot}} = S^2_A + S^2_B + S^2_{\text{Int}} + S^2_{\text{Resid}}, \tag{11.8.9}$$

where

$$S_A^2 = JK \sum_{i=1}^{I} (\overline{Y}_{i++} - \overline{Y}_{+++})^2, \tag{11.8.10}$$

$$S_B^2 = IK \sum_{j=1}^{J} (\overline{Y}_{+j+} - \overline{Y}_{+++})^2,$$

$$S_{\text{Int}}^2 = K \sum_{i=1}^{I} \sum_{j=1}^{J} (\overline{Y}_{ij+} - \overline{Y}_{i++} - \overline{Y}_{+j+} + \overline{Y}_{+++})^2,$$

$$S_{\text{Resid}}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (Y_{ijk} - \overline{Y}_{ij+})^2.$$

In addition, $S_{\text{Resid}}^2/\sigma^2$ has the χ^2 distribution with $IJ(K-1)$ degrees of freedom. If all $\alpha_i = 0$, then S_A^2/σ^2 has the χ^2 distribution with $I-1$ degrees of freedom. If all $\beta_j = 0$, then S_B^2/σ^2 has the χ^2 distribution with $J-1$ degrees of freedom. If all $\gamma_{ij} = 0$, then $S_{\text{Int}}^2/\sigma^2$ has the χ^2 distribution with $(I-1)(J-1)$ degrees of freedom. The four sums of squares are mutually independent.

Proof The proof of (11.8.9) is left to the reader in Exercise 7.

The random variable $S_{\text{Resid}}^2/\sigma^2$ is the sum of IJ independent random variables of the form $\sum_{k=1}^{K}(Y_{ijk} - \overline{Y}_{ij+})^2/\sigma^2$. According to Theorem 8.3.1, each of these IJ random variables has the χ^2 distribution with $K-1$ degrees of freedom. Hence, the sum of all IJ of them has the χ^2 distribution with $IJ(K-1)$ degrees of freedom. If all of the $\alpha_i = 0$, then $\overline{Y}_{1++}, \dots, \overline{Y}_{I++}$ all have the normal distribution with mean μ and variance σ^2/JK. Theorem 8.3.1 implies that S_A^2/σ^2 has the χ^2 distribution with $I-1$ degrees of freedom. Similarly, if all $\beta_j = 0$, then S_B^2/σ^2 has the χ^2 distribution with $J-1$ degrees of freedom.

The number of degrees of freedom for S_{Int}^2 can be determined as follows: If all of the $\gamma_{ij} = 0$, then the additivity assumption holds, and S_{Int}^2 is the same as S_{Resid}^2 from Sec. 11.7 except for the fact that each $\overline{Y}_{ij+}$ has the normal distribution with mean $\mu + \alpha_i + \beta_j$ and variance σ^2/K instead of variance σ^2. This means that if all $\gamma_{ij} = 0$, then $S_{\text{Int}}^2/\sigma^2$ has the χ^2 distribution with $(I-1)(J-1)$ degrees of freedom.

Finally, it can be shown that all of the sums of squares in relations (11.8.10) are independent (see Exercise 8 for a related result). ■

The claims in Theorem 11.8.2 are summarized in Table 11.28, which is the ANOVA table for the two-way layout with K observations per cell.

Example 11.8.3

Gasoline Consumption. Using the sample means computed in Example 11.8.2, we can form the ANOVA table in Table 11.29. We shall use the mean squares in Table 11.29 to test various hypotheses about the effects of the factors after we develop test procedures. ◄

Testing Hypotheses

As mentioned before, the effects of the factors A and B are additive if and only if all the interactions γ_{ij} vanish. Hence, to test whether the effects of the factors are

Table 11.28 General ANOVA table for two-way layout with replication

Source of variation	Degrees of freedom	Sum of squares	Mean square
Main effects of A	$I-1$	S_A^2	$S_A^2/(I-1)$
Main effects of B	$J-1$	S_B^2	$S_B^2/(J-1)$
Interactions	$(I-1)(J-1)$	S_{Int}^2	$S_{\text{Int}}^2/[(I-1)(J-1)]$
Residuals	$IJ(K-1)$	S_{Resid}^2	$S_{\text{Resid}}^2/[IJ(K-1)]$
Total	$IJK-1$	S_{Tot}^2	

Table 11.29 ANOVA table for data from Example 11.8.2.

Source of variation	Degrees of freedom	Sum of squares	Mean square
Main effects of device	1	0.4813	0.4813
Main effects of model	2	30.92	15.46
Interactions	2	0.1147	0.0573
Residuals	24	18.22	0.7590
Total	29	49.73	

additive, we must test the following hypotheses:

$$\begin{aligned} H_0: &\quad \gamma_{ij}=0 \quad \text{for } i=1,\ldots,I \text{ and } j=1,\ldots,J,\\ H_1: &\quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{11.8.11}$$

It follows from Theorem 11.8.2 that when the null hypothesis H_0 is true, the random variable $S_{\text{Int}}^2/\sigma^2$ has the χ^2 distribution with $(I-1)(J-1)$ degrees of freedom. Furthermore, regardless of whether or not H_0 is true, the independent random variable $S_{\text{Resid}}^2/\sigma^2$ has the χ^2 distribution with $IJ(K-1)$ degrees of freedom. Thus, when H_0 is true, the following random variable U_{AB}^2 has the F distribution with $(I-1)(J-1)$ and $IJ(K-1)$ degrees of freedom:

$$U_{AB}^2 = \frac{IJ(K-1)S_{\text{Int}}^2}{(I-1)(J-1)S_{\text{Resid}}^2}, \tag{11.8.12}$$

which is also the ratio of the interaction mean square to the residual mean square.

The null hypothesis H_0 would be rejected at level α_0 if

$$U_{AB}^2 \ge F_{(I-1)(J-1),IJ(K-1)}^{-1}(1-\alpha_0),$$

where $F_{(I-1)(J-1),IJ(K-1)}^{-1}$ is the quantile function of the F distribution with $(I-1)(J-1)$ and $IJ(K-1)$ degrees of freedom.

Example 11.8.4 Gasoline Consumption. Suppose that it is desired to use the data from Example 11.8.2 to test the null hypothesis that the effects of equipping a car with the device and using a particular model are additive, against the alternative that these effects are not additive. In other words, suppose that it is desired to test the hypotheses (11.8.11). Using the mean squares in Table 11.29 and Eq. (11.8.12), we compute that $U^2_{AB} = 0.0573/0.7590 = 0.076$. The corresponding p-value can be found using statistical software, and its value is 0.9275. Hence, the null hypothesis that the effects are additive would be not be rejected at any common level of significance. ◀

If the null hypothesis H_0 in (11.8.11) is rejected, then it suggests that at least some of the interactions γ_{ij} are not 0. Therefore, the means of the observations for certain combinations of i and j will be larger than the means of the observations for other combinations, and both factor A and factor B affect these means. In this case, because both factor A and factor B affect the means of the observations, there is not usually any further interest in testing whether either the main effects $\alpha_1, \dots, \alpha_I$ or the main effects $\beta_1, \dots, \beta_J$ are zero.

On the other hand, if the null hypothesis H_0 in (11.8.11) is not rejected (as is the case in Example 11.8.4), then we might decide to act as if all the interactions are 0. If, in addition, all the main effects $\alpha_1, \dots, \alpha_I$ were 0, then the mean value of each observation would not depend in any way on the value of i. In this case, factor A would have no effect on the observations. Therefore, if the null hypothesis H_0 in (11.8.11) is not rejected, we might be interested in testing the following hypotheses:

$$\begin{aligned} H_0:&\quad \alpha_i = 0 \text{ and } \gamma_{ij} = 0 \text{ for } i = 1, \dots, I \text{ and } j = 1, \dots, J, \\ H_1:&\quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{11.8.13}$$

Indeed, we might be interested in testing these hypotheses even if we had not first tested the hypotheses (11.8.11).

According to Theorem 11.8.2, if H_0 is true, then S^2_A/σ^2 and $S^2_{\text{Int}}/\sigma^2$ are independent having χ^2 distributions with $I-1$ and $(I-1)(J-1)$ degrees of freedom, respectively. It follows that, when H_0 in (11.8.13) is true, the following random variable U^2_A has the F distribution with $I-1+(I-1)(J-1) = (I-1)J$ and $IJ(K-1)$ degrees of freedom:

$$U^2_A = \frac{IJ(K-1)[S^2_A + S^2_{\text{Int}}]}{(I-1)J S^2_{\text{Resid}}}. \tag{11.8.14}$$

If we did not test the hypotheses (11.8.11) first, then we can reject H_0 in (11.8.13) at level α_0 if $U^2_A \ge F^{-1}_{(I-1)J,IJ(K-1)}(1-\alpha_0)$.

If we first tested (11.8.11) and failed to reject the null hypothesis, there are two important considerations to emphasize before proceeding with a test of (11.8.13). First, the size of the second test, the test of (11.8.13), should be calculated *conditional* on having failed to reject the null hypothesis in (11.8.11). That is, if the second test is to reject the null hypothesis in (11.8.13) if $T \ge c$ (for some statistic T), then the size of the second test should be the conditional probability

$$\Pr\Big(T \ge c \;\Big|\; U^2_{AB} < F^{-1}_{(I-1)(J-1),IJ(K-1)}(1-\alpha_0)\Big). \tag{11.8.15}$$

Calculation of this conditional probability is beyond the scope of this book, but it can be approximated using simulation methods that will be introduced in Chapter 12. (See Example 12.3.4 for an illustration.)

The second consideration involves the choice of test statistic T for testing (11.8.13). For the case in which we did not first test (11.8.11), the statistic U^2_A in

(11.8.14) is a sensible test statistic. However, if we have already failed to reject the null hypothesis in (11.8.11), a better test statistic might be

$$V_A^2 = \frac{IJ(K-1)S_A^2}{(I-1)S_{\text{Resid}}^2}. \tag{11.8.16}$$

One reason for this is that, with $T = V_A^2$, the probability in (11.8.15) will often be closer to α_0 than with $T = U_A^2$. For instance, if $IJ(K-1)$ is large and H_0 is true, then S_{Resid}^2 should be close to σ^2 with high probability. In this case, since S_{Int}^2 and S_A^2 are independent random variables, the random variables V_A^2 and U_{AB}^2 should be nearly independent as well. This will make the test based on V_A^2 nearly independent of whether or not the test based on U_{AB}^2 rejected its null hypothesis. On the other hand, because

$$U_A^2 = \frac{1}{J}[V_A^2 + (J-1)U_{AB}^2],$$

we see that the dependence between U_A^2 and U_{AB}^2 is likely to be quite high under all circumstances.

So, if we first test (11.8.11) and fail to reject the null hypothesis, we should then use V_A^2 to test (11.8.13). We would then reject the null hypothesis if $V_A^2 > c$, where c is some constant. Unfortunately, we still cannot find a useable expression for c other than to note that the size of this second test, conditional on the first test, is (11.8.15) with $T = V_A^2$. We can use simulation methods to compute this if necessary. (See Example 12.3.4.) The overall size of this two-stage procedure is larger than α_0. (See Exercise 19.) In practice, it is common to let $c = F^{-1}_{I-1,IJ(K-1)}(1-\alpha_0)$ and pretend as if (11.8.15) with $T = V_A^2$ is essentially α_0.

Example 11.8.5

Gasoline Consumption. Suppose now that it is desired to test the null hypothesis that the device has no effect on gasoline consumption for all of the car models tested, against the alternative that the device does affect gasoline consumption. In other words, suppose that it is desired to test the hypotheses (11.8.13). If we had not first tested (11.8.11), then we would use Eq. (11.8.14) and the numbers in Table 11.29 to compute $U_A^2 = 24(0.4813 + 0.1147)/[3(18.22)] = 0.2616$. The corresponding p-value from the F distribution with 3 and 24 degrees of freedom is 0.8523. Hence, the null hypothesis would not be rejected at the usual levels of significance.

On the other hand, since we did test (11.8.11) first, we should instead use $V_A^2 = 0.4813/0.7590 = 0.6341$. We cannot compute the exact conditional p-value associated with this observed value. However, using the method to be described in Example 12.3.4, we can approximate the p-value to be about 0.43, given that we failed to reject the null hypothesis in (11.8.11). We can also use the method of Example 12.3.4 to approximate the probabilities in (11.8.15) for $T = U_A^2$ and for $T = V_A^2$. With $\alpha_0 = 0.05$, these approximations are, respectively, 0.019 and 0.048. Notice how close the test based on V_A^2 comes to having the nominal size $\alpha_0 = 0.05$, while the conditional size of the test based on U_A^2 is much smaller. ◀

Similarly, we may want to find out whether all the main effects of factor B, as well as the interactions, are 0. In this case, we would test the following hypotheses:

$$\begin{aligned} H_0&: \quad \beta_j = 0 \text{ and } \gamma_{ij} = 0 \quad \text{for } i = 1, \dots, I, \text{ and } j = 1, \dots, J, \\ H_1&: \quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned} \tag{11.8.17}$$

By analogy with Eq. (11.8.14), it follows that when H_0 is true, the following random variable U_B^2 has the F distribution with $I(J-1)$ and $IJ(K-1)$ degrees of freedom:

$$U_B^2 = \frac{IJ(K-1)[S_B^2 + S_{\text{Int}}^2]}{I(J-1)S_{\text{Resid}}^2}. \tag{11.8.18}$$

Again, if we do not first test (11.8.11), then the hypothesis H_0 should be rejected at level α_0 if $U_B^2 > F^{-1}_{I(J-1),IJ(K-1)}(1-\alpha_0)$. If we test (11.8.11) first and fail to reject the null hypothesis, then we should reject H_0 in (11.8.17) if V_B^2 is too large, where $V_B^2 = \frac{IJ(K-1)S_B^2}{(J-1)S_{\text{Resid}}^2}$. The conditional level of this test must be computed by simulation, also.

In a given problem, if the null hypothesis in (11.8.11) is not rejected and the null hypotheses in both (11.8.13) and (11.8.17) are rejected, then we may be willing to proceed with further studies and experimentation by using a model in which it is assumed that the effects of factor A and factor B are approximately additive and the effects of both factors are important.

The results obtained in Example 11.8.5 do not provide any indication that the device is effective. Nevertheless, it can be seen from Table 11.27 that for each of the three models, the average consumption of gasoline for the cars that were equipped with the device is smaller than the average consumption for the cars that were not so equipped. If we assume that the effects of the device and the model of automobile are additive, then regardless of the model of the automobile that is used, the M.L.E. of the reduction in gasoline consumption over the given route that is achieved by equipping an automobile with the device is $\hat{\alpha}_2 - \hat{\alpha}_1 = 0.2534$ liter.

The Two-Way Layout with Unequal Numbers of Observations in the Cells

Consider again a two-way layout with I rows and J columns, but suppose now that instead of there being K observations in each cell, some cells have more observations than others. For $i = 1, \ldots, I$ and $j = 1, \ldots, J$, we shall let K_{ij} denote the number of observations in the (i, j) cell. Thus, the total number of observations is $\sum_{i=1}^{I}\sum_{j=1}^{J} K_{ij}$. We shall assume that every cell contains at least one observation, and we shall again let Y_{ijk} denote the kth observation in the (i, j) cell. For each value of i and j, the values of the subscript k are $1, \ldots, K_{ij}$. We shall also assume, as before, that all the observations Y_{ijk} are independent; each has a normal distribution; $\text{Var}(Y_{ijk}) = \sigma^2$ for all values of i, j, and k; and $E(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$, where these parameters satisfy the conditions given in Eq. (11.8.5).

As usual, we shall let $\overline{Y}_{ij+}$ denote the average of the observations in the (i, j) cell. It can then be shown that for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, the M.L.E.'s, or least-squares estimators, are as follows:

$$\begin{aligned}
\hat{\mu} &= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\overline{Y}_{ij+}, \quad \hat{\alpha}_i = \frac{1}{J}\sum_{j=1}^{J}\overline{Y}_{ij+} - \hat{\mu}, \\
\hat{\beta}_j &= \frac{1}{I}\sum_{i=1}^{I}\overline{Y}_{ij+} - \hat{\mu}, \quad \hat{\gamma}_{ij} = \overline{Y}_{ij+} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j.
\end{aligned} \tag{11.8.19}$$

These estimators are intuitively reasonable and analogous to those given in Eqs. (11.8.6) and (11.8.7).

Suppose now, however, that it is desired to test hypotheses such as (11.8.11), (11.8.13), or (11.8.17). The construction of appropriate tests becomes somewhat more difficult because, in general, the sums of squares analogous to those given in Eq. (11.8.10) will not be independent when there are unequal numbers of observations in the different cells. Hence, the test procedures presented earlier in this section cannot be directly copied here. It is necessary to develop other sums of squares that will be independent and will reflect the different types of variations in the data in which we are interested. We shall not consider the problem further in this book. This and other problems of ANOVA are described in the advanced book by Scheffé (1959).

Summary

We extended the analysis of the two-way layout to cases in which we have equal numbers of observations at all combinations of levels of the two factors. One additional null hypothesis that we can test in this case is that the effects of the two factors are additive. (We assumed that the effects were additive when we had only one observation per cell.) If we reject the null hypothesis of additivity, we typically do not test any further hypotheses. If we don't reject this null hypothesis, we might still be interested in whether one of the two factors has any effect at all on the means of the observations. Even if we do not first test the null hypothesis that the effects of the two factors are additive, we might still be interested in whether one of the factors has an effect. The precise form of a test of one of these last hypotheses depends on whether we first test that the effects are additive.

Exercises

1. Show that for every set of numbers θ_{ij} $(i = 1, \dots, I$ and $j = 1, \dots, J)$, there exists a unique set of numbers μ, α_i, β_j, and γ_{ij} $(i = 1, \dots, I$ and $j = 1, \dots, J)$ that satisfy Eqs. (11.8.4) and (11.8.5).

2. Verify that Eq. (11.8.6) gives the M.L.E.'s of the parameters of the two-way layout with replication.

3. Suppose that in a two-way layout, the values of θ_{ij} are as given in each of the four matrices presented in parts (a), (b), (c), and (d) of Exercise 2 of Sec. 11.7. For each matrix, determine the values of μ, α_i, β_j, and γ_{ij} that satisfy Eqs. (11.8.4) and (11.8.5).

4. Verify that if $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$ are as given by Eqs. (11.8.6) and (11.8.7), then $\sum_{i=1}^{I} \hat{\alpha}_i = 0$, $\sum_{j=1}^{J} \hat{\beta}_j = 0$, $\sum_{i=1}^{I} \hat{\gamma}_{ij} = 0$ for $j = 1, \dots, J$, and $\sum_{j=1}^{J} \hat{\gamma}_{ij} = 0$ for $i = 1, \dots, I$.

5. Verify that if $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$ are as given by Eqs. (11.8.6) and (11.8.7), then $E(\hat{\mu}) = \mu$, $E(\hat{\alpha}_i) = \alpha_i$, $E(\hat{\beta}_j) = \beta_j$, and $E(\hat{\gamma}_{ij}) = \gamma_{ij}$ for all values of i and j. *Hint:* Each of the random variables in this exercise is a linear function of the Y_{ijk}'s, and hence the expectations are the same linear combinations of the expectations of the Y_{ijk}'s.

6. Show that if $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$ are as given by Eqs. (11.8.6) and (11.8.7), then the following results are true for all values of i and j:

$$\text{Var}(\hat{\mu}) = \frac{I}{IJK}\sigma^2, \qquad \text{Var}(\hat{\alpha}_i) = \frac{(I-1)}{IJK}\sigma^2,$$

$$\text{Var}(\hat{\beta}_j) = \frac{(J-1)}{IJK}\sigma^2, \qquad \text{Var}(\hat{\gamma}_{ij}) = \frac{(I-1)(J-1)}{IJK}\sigma^2.$$

7. Verify Eq. (11.8.9).

8. In a two-way layout with K observations in each cell, show that for all values of i, i_1, i_2, j, j_1, j_2, and k, the following five random variables are uncorrelated with one another:

$$Y_{ijk} - \overline{Y}_{ij+},\ \hat{\alpha}_{i_1},\ \hat{\beta}_{j_1},\ \hat{\gamma}_{i_2 j_2},\ \text{and } \hat{\mu}.$$

9. Verify that U^2_{AB} also equals

$$\frac{\left(\begin{array}{c} IJK(K-1)(\sum_{i=1}^{I}\sum_{j=1}^{J}\overline{Y}^2_{ij+} - J\sum_{i=1}^{I}\overline{Y}^2_{i++} \\ -I\sum_{j=1}^{J}\overline{Y}^2_{+j+} + IJ\overline{Y}^2_{+++}) \end{array}\right)}{\left(\begin{array}{c} (I-1)(J-1)(\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}Y^2_{ijk} \\ -K\sum_{i=1}^{I}\sum_{j=1}^{J}\overline{Y}^2_{ij+}) \end{array}\right)}.$$

10. Suppose that in an experimental study to determine the combined effects of receiving both a stimulant and a tranquilizer, three different types of stimulants and four different types of tranquilizers are administered to a group of rabbits. Each rabbit in the experiment receives one of the stimulants and then, 20 minutes later, receives one of the tranquilizers. After one hour, the response of the rabbit is measured in appropriate units. In order that each possible pair of drugs may be administered to two different rabbits, 24 rabbits are used in the experiment. The responses of these 24 rabbits are given in Table 11.30. Determine the values of $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$ for $i = 1, 2, 3$ and $j = 1, 2, 3, 4$, and determine also the value of $\hat{\sigma}^2$.

Table 11.30 Data for Exercises 10–15

	Tranquilizer			
Stimulant	1	2	3	4
1	11.2	7.4	7.1	9.6
	11.6	8.1	7.0	7.6
2	12.7	10.3	8.8	11.3
	14.0	7.9	8.5	10.8
3	10.1	5.5	5.0	6.5
	9.6	6.9	7.3	5.7

11. For the conditions of Exercise 10 and the data in Table 11.30, test the hypothesis that every interaction between a stimulant and a tranquilizer is 0.

12. For the conditions of Exercise 10 and the data in Table 11.30, test the hypothesis that all three stimulants yield the same responses.

13. For the conditions of Exercise 10 and the data in Table 11.30, test the hypothesis that all four tranquilizers yield the same responses.

14. For the conditions of Exercise 10 and the data in Table 11.30, test the following hypotheses:

$$H_0: \quad \mu = 8,$$
$$H_1: \quad \mu \neq 8.$$

15. For the conditions of Exercise 10 and the data in Table 11.30, test the following hypotheses:

$$H_0: \quad \alpha_2 \le 1,$$
$$H_1: \quad \alpha_2 > 1.$$

16. In a two-way layout with unequal numbers of observations in the cells, show that if $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$ are as given by Eq. (11.8.19), then $E(\hat{\mu}) = \mu$, $E(\hat{\alpha}_i) = \alpha_i$, $E(\hat{\beta}_j) = \beta_j$, and $E(\hat{\gamma}_{ij}) = \gamma_{ij}$ for all values of i and j.

17. Verify that if $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$ are as given by Eq. (11.8.19), then $\sum_{i=1}^{I}\hat{\alpha}_i = 0$, $\sum_{j=1}^{J}\hat{\beta}_j = 0$, $\sum_{i=1}^{I}\hat{\gamma}_{ij} = 0$ for $j = 1, \dots, J$, and $\sum_{j=1}^{J}\hat{\gamma}_{ij} = 0$ for $i = 1, \dots, I$.

18. Show that if $\hat{\mu}$ and $\hat{\alpha}_i$ are as given by Eq. (11.8.19), then for $i = 1, \dots, I$,

$$\text{Cov}(\hat{\mu}, \hat{\alpha}_i) = \frac{\sigma^2}{IJ^2}\left[\sum_{j=1}^{J}\frac{1}{K_{ij}} - \frac{1}{I}\sum_{r=1}^{I}\sum_{j=1}^{J}\frac{1}{K_{rj}}\right].$$

Also, show that this covariance is 0 if all K_{ij} are the same.

19. Recall the two-stage testing procedure described in this section: First test (11.8.11) at level α_0. If you reject the null hypothesis, stop. If you don't reject the null hypothesis, then test (11.8.13). Let β_0 be the conditional size of the second test given that the first test doesn't reject the null hypothesis. Assume that both null hypotheses are true. Find the probability that this two-stage procedure rejects at least one of the null hypotheses.

20. The study referred to in Exercise 10 in Sec. 11.6 actually included another factor in addition to size of vehicle. There were two different filters, a standard filter and a newly developed filter. Table 11.19 has data only from the standard filter. The corresponding data for the new filter are in Table 11.31.

Table 11.31 Data for Exercise 20. This table has data for the vehicles with the new filter.

Vehicle size	Noise values
Small	820, 820, 820, 825, 825, 825
Medium	820, 820, 825, 815, 825, 825
Large	775, 775, 775, 770, 760, 765

a. Construct the ANOVA table for the two-way layout that includes the data from both Tables 11.19 and 11.31.

b. Compute the p-value for testing the null hypothesis that there is no interaction.

c. Compute the p-value for testing the null hypothesis that the vehicles of all three sizes produce the same level of noise on average.

d. Compute the p-value for testing the null hypothesis that both filters result in the same level of noise on average.

11.9 Supplementary Exercises

1. Consider the data in Example 11.2.2 on page 703. Suppose that we fit a simple linear regression of the natural logarithm of pressure on boiling point.

a. Find a 90 percent confidence interval (bounded interval) for the slope β_1.

b. Test the null hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ at level $\alpha_0 = 0.1$.

c. Find a 90 percent prediction interval for pressure (not logarithm of pressure) when the boiling point is 204.6.

2. Suppose that (X_i, Y_i), $i = 1, \dots, n$, form a random sample of size n from the bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ, and let $\hat{\mu}_i$, $\hat{\sigma}_i^2$, and $\hat{\rho}$ denote their M.L.E.'s Also, let $\hat{\beta}_2$ denote the M.L.E. of β_2 in the regression of Y on X. Show that

$$\hat{\beta}_2 = \hat{\rho}\hat{\sigma}_2/\hat{\sigma}_1.$$

Hint: See Exercise 24 of Sec. 7.6.

3. Suppose that (X_i, Y_i), $i = 1, \dots, n$, form a random sample of size n from the bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ. Determine the mean and the variance of the following statistic T, given the observed values $X_1 = x_1, \dots, X_n = x_n$:

$$T = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

4. Let θ_1, θ_2, and θ_3 denote the unknown angles of a triangle, measured in degrees ($\theta_i > 0$ for $i = 1, 2, 3$, and $\theta_1 + \theta_2 + \theta_3 = 180$). Suppose that each angle is measured by an instrument that is subject to error, and the measured values of θ_1, θ_2, and θ_3 are found to be $y_1 = 83$, $y_2 = 47$, and $y_3 = 56$, respectively. Determine the least-squares estimates of θ_1, θ_2, and θ_3.

5. Suppose that a straight line is to be fitted to n points $(x_1, y_1), \dots, (x_n, y_n)$ such that $x_2 = x_3 = \cdots = x_n$ but $x_1 \neq x_2$. Show that the least-squares line will pass through the point (x_1, y_1).

6. Suppose that a least-squares line is fitted to the n points $(x_1, y_1), \dots, (x_n, y_n)$ in the usual way by minimizing the sum of squares of the vertical deviations of the points from the line, and another least-squares line is fitted by minimizing the sum of squares of the horizontal deviations of the points from the line. Under what conditions will these two lines coincide?

7. Suppose that a straight line $y = \beta_1 + \beta_2 x$ is to be fitted to the n points $(x_1, y_1), \dots, (x_n, y_n)$ in such a way that the sum of the squared perpendicular (or orthogonal) distances from the points to the line is a minimum. Determine the optimal values of β_1 and β_2.

8. Suppose that twin sisters are each to take a certain mathematics examination. They know that the scores they will obtain on the examination have the same mean μ, the same variance σ^2, and positive correlation ρ. Assuming that their scores have a bivariate normal distribution, show that after each twin learns her own score, the expected value of her sister's score is between her own score and μ.

9. Suppose that a sample of n observations is formed from k subsamples containing $n_1, \dots, n_k$ observations ($n_1 + \cdots + n_k = n$). Let x_{ij} ($j = 1, \dots, n_i$) denote the observations in the ith subsample, and let $\overline{x}_{i+}$ and v_i^2 denote the sample mean and the sample variance of that subsample:

$$\bar{x}_{i+} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \quad v_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{i+})^2.$$

Finally, let $\overline{x}_{++}$ and v^2 denote the sample mean and the sample variance of the entire sample of n observations:

$$\overline{x}_{++} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}, \quad v^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{++})^2.$$

Determine an expression for v^2 in terms of $\overline{x}_{++}$, $\overline{x}_{i+}$, and v_i^2 ($i = 1, \dots, k$).

10. Consider the linear regression model

$$Y_i = \beta_1 w_i + \beta_2 x_i + \varepsilon_i \quad \text{for } i = 1, \dots, n,$$

where $(w_1, x_1), \ldots, (w_n, x_n)$ are given pairs of constants and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random variables, each of which has the normal distribution with mean 0 and variance σ^2. Determine explicitly the M.L.E.'s of β_1 and β_2.

11. Determine an unbiased estimator of σ^2 in a two-way layout with K observations in each cell ($K \geq 2$).

12. In a two-way layout with one observation in each cell, construct a test of the null hypothesis that all the effects of both factor A and factor B are 0.

13. In a two-way layout with K observations in each cell ($K \geq 2$), construct a test of the null hypothesis that all the main effects for factor A and factor B, and also all the interactions, are 0.

14. Suppose that each of two different varieties of corn is treated with two different types of fertilizer in order to compare the yields, and that K independent replications are obtained for each of the four combinations. Let X_{ijk} denote the yield on the kth replication of the combination of variety i with fertilizer j ($i = 1, 2;\ j = 1, 2;\ k = 1, \ldots, K$). Assume that all the observations are independent and normally distributed, each distribution has the same unknown variance, and $E(X_{ijk}) = \mu_{ij}$ for $k = 1, \ldots, K$. Explain in words what the following hypotheses mean, and describe how to carry out a test of them:

$$H_0: \quad \mu_{11} - \mu_{12} = \mu_{21} - \mu_{22},$$
$$H_1: \quad \text{The hypothesis } H_0 \text{ is not true.}$$

15. Suppose that W_1, W_2, and W_3 are independent random variables, each of which has a normal distribution with the following means and variances:

$$E(W_1) = \theta_1 + \theta_2, \qquad \text{Var}(W_1) = \sigma^2,$$
$$E(W_2) = \theta_1 + \theta_2 - 5, \ \text{Var}(W_2) = \sigma^2,$$
$$E(W_3) = 2\theta_1 - 2\theta_2, \quad \text{Var}(W_3) = 4\sigma^2.$$

Determine the M.L.E.'s of θ_1, θ_2, and σ^2, and determine also the joint distribution of these estimators.

16. Suppose that it is desired to fit a curve of the form $y = \alpha x^{\beta}$ to a given set of n points (x_i, y_i) with $x_i > 0$ and $y_i > 0$ for $i = 1, \ldots, n$. Explain how this curve can be fitted either by direct application of the method of least squares or by first transforming the problem into one of fitting a straight line to the n points $(\log x_i, \log y_i)$ and then applying the method of least squares. Discuss the conditions under which each of these methods is appropriate.

17. Consider a problem of simple linear regression, and let $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ denote the residual of the observation Y_i ($i = 1, \ldots, n$), as defined in Definition 11.3.2. Evaluate $\text{Var}(e_i)$ for given values of $x_1, \ldots, x_n$, and show that it is a decreasing function of the distance between x_i and $\overline{x}$.

18. Consider a general linear model with $n \times p$ design matrix $\boldsymbol{Z}$, and let $\boldsymbol{W} = \boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}$ denote the vector of residuals. (In other words, the ith coordinate of $\boldsymbol{W}$ is $Y_i - \hat{Y}_i$, where $\hat{Y}_i = z_{i0}\hat{\beta}_0 + \cdots + z_{ip-1}\hat{\beta}_{p-1}$.

a. Show that $\boldsymbol{W} = \boldsymbol{D}\boldsymbol{Y}$, where

$$\boldsymbol{D} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'.$$

b. Show that the matrix $\boldsymbol{D}$ is *idempotent*; that is, $\boldsymbol{D}\boldsymbol{D} = \boldsymbol{D}$.

c. Show that $\text{Cov}(\boldsymbol{W}) = \sigma^2\boldsymbol{D}$.

19. Consider a two-way layout in which the effects of the factors are additive so that Eq. (11.7.1) is satisfied, and let $v_1, \ldots, v_I$ and $w_1, \ldots, w_J$ be arbitrary given positive numbers. Show that there exist unique numbers $\mu, \alpha_1, \ldots, \alpha_I$, and $\beta_1, \ldots, \beta_J$ such that

$$\sum_{i=1}^{I} v_i\alpha_i = \sum_{j=1}^{J} w_j\beta_j = 0$$

and

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j \quad \text{for } i = 1, \ldots, I \text{ and } j = 1, \ldots, J.$$

20. Consider a two-way layout in which the effects of the factors are additive, as in Exercise 19; suppose also that there are K_{ij} observations per cell, where $K_{ij} > 0$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Let $v_i = K_{i+}$ for $i = 1, \ldots, I$, and let $w_j = K_{+j}$ for $j = 1, \ldots, J$. Assume that $E(Y_{ijk}) = \mu + \alpha_i + \beta_j$ for $k = 1, \ldots, K_{ij}$, $i = 1, \ldots, j$, and $j = 1, \ldots, J$, where $\sum_{i=1}^{I} v_i\alpha_i = \sum_{j=1}^{J} w_j\beta_j = 0$, as in Exercise 19. Verify that the least-squares estimators of μ, α_i, and β_j are as follows:

$$\hat{\mu} = \overline{Y}_{+++},$$
$$\hat{\alpha}_i = \frac{1}{K_{i+}}Y_{i++} - \overline{Y}_{+++} \quad \text{for } i = 1, \ldots, I,$$
$$\hat{\beta}_j = \frac{1}{K_{+j}}Y_{+j+} - \overline{Y}_{+++} \quad \text{for } j = 1, \ldots, J.$$

21. Consider again the conditions of Exercises 19 and 20, and let the estimators $\hat{\mu}$, $\hat{\alpha}_i$, and $\hat{\beta}_j$ be as given in Exercise 20. Show that $\text{Cov}(\hat{\mu}, \hat{\alpha}_i) = \text{Cov}(\hat{\mu}, \hat{\beta}_j) = 0$.

22. Consider again the conditions of Exercise 19 and 20, and suppose that the numbers K_{ij} have the following proportionality property:

$$K_{ij} = \frac{K_{i+}K_{+j}}{n} \quad \text{for } i = 1, \ldots, I \text{ and } j = 1, \ldots, J.$$

Show that $\text{Cov}(\hat{\alpha}_i, \hat{\beta}_j) = 0$, where the estimators $\hat{\alpha}_i$ and $\hat{\beta}_j$ are as given in Exercise 20.

23. In a three-way layout with one observation in each cell, the observations Y_{ijk} $(i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, K)$ are assumed to be independent and normally distributed, with a common variance σ^2. Suppose that $E(Y_{ijk}) = \theta_{ijk}$. Show that for every set of numbers θ_{ijk}, there exists a unique set of numbers μ, α_i^A, α_j^B, α_k^C, β_{ij}^{AB}, β_{ik}^{AC}, β_{jk}^{BC}, and γ_{ijk} $(i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, K)$ such that

$$\alpha_+^A = \alpha_+^B = \alpha_+^C = 0,$$
$$\beta_{i+}^{AB} = \beta_{+j}^{AB} = \beta_{i+}^{AC} = \beta_{+k}^{AC} = \beta_{j+}^{BC} = \beta_{+k}^{BC} = 0,$$
$$\gamma_{ij+} = \gamma_{i+k} = \gamma_{+jk} = 0,$$

and

$$\theta_{ijk} = \mu + \alpha_i^A + \alpha_j^B + \alpha_k^C + \beta_{ij}^{AB} + \beta_{ik}^{AC} + \beta_{jk}^{BC} + \gamma_{ijk},$$

for all values of i, j, and k.

24. The 2000 U.S. presidential election was very close, especially in the state of Florida. Indeed, newscasters were unable to predict a winner the day after the election because they could not decide who was going to win Florida's 25 electoral votes. Many voters in Palm Beach County complained that they were confused by the design of the ballot and might have voted for Patrick Buchanan instead of Al Gore, as they had intended. Table 11.32 displays the official ballot counts (after all official recounts) for each county. There was no reason, prior to the election, to think that Patrick Buchanan would gather a significantly higher percent of the vote in Palm Beach County than in any other Florida county.

a. Draw a plot of the vote count for Patrick Buchanan against the total vote count with one point for each county. Identify the point corresponding to Palm Beach County.

b. Given the complaints about the Palm Beach ballot, it might be sensible to treat the data point for Palm Beach County as being different from the others. Fit a simple linear regression model with Y being the vote for Buchanan and X being the total vote in each county, excluding Palm Beach County.

c. Plot the residuals from the regression in part (b) against the X variable. Do you notice any pattern in the plot?

d. The variance of the vote for each candidate in a county ought to depend on the total vote in the county. The larger the total vote, the more variance you expect in the vote for each candidate. For this reason, the assumptions of the simple linear regression model would not hold. As an alternative, fit a simple linear regression with Y being the logarithm of the vote for Buchanan and X being the logarithm of the total vote in each county. Continue to exclude Palm Beach County.

e. Plot the residuals from the regression in part (d) against the X variable. Do you notice any pattern in the plot?

f. Using the model fit in part (d), form a 99 percent prediction interval for the Buchanan vote (not the logarithm of the Buchanan vote) in Palm Beach County.

g. Let B be the upper endpoint of the interval you found in part (f). Just suppose that the actual number of people in Palm Beach County who voted for Buchanan had actually been B instead of 3411. Also suppose that the remaining $3411 - B$ voters had actually voted for Gore. Would this have changed the winner of the total popular vote for the State of Florida?

Table 11.32 County votes for Bush, Gore, and Buchanan in the 2000 presidential election for the state of Florida. The total column includes all 11 candidates that were on the ballot. The absentee row includes overseas absentee ballots that were not included in individual county totals. These data came from the official state of Florida election Web site, which has since been moved or deleted.

County	Bush	Gore	Buchanan	Total	County	Bush	Gore	Buchanan	Total
Alachua	34,124	47,365	263	85,729	Lee	106,141	73,560	305	184,377
Baker	5610	2392	73	8154	Leon	39,062	61,427	282	103,124
Bay	38,637	18,850	248	58,805	Levy	6858	5398	67	12,724
Bradford	5414	3075	65	8673	Liberty	1317	1017	39	2410
Brevard	115,185	97,318	570	218,395	Madison	3038	3014	29	6162
Broward	177,902	387,703	795	575,143	Manatee	57,952	49,177	271	110,221
Calhoun	2873	2155	90	5174	Marion	55,141	44,665	563	102,956
Charlotte	35,426	29,645	182	66,896	Martin	33,970	26,620	112	62,013
Citrus	29,767	25,525	270	57,204	Miami-Dade	289,533	328,808	560	625,449
Clay	41,736	14,632	186	57,353	Monroe	16,059	16,483	47	33,887
Collier	60,450	29,921	122	92,162	Nassau	16,404	6952	90	23,780
Columbia	10,964	7047	89	18,508	Okaloosa	52,093	16,948	267	70,680
Desoto	4256	3320	36	7811	Okeechobee	5057	4588	43	9853
Dixie	2697	1826	29	4666	Orange	134,517	140,220	446	280,125
Duval	152,098	107,864	652	264,636	Osceola	26,212	28,181	145	55,658
Escambia	73,017	40,943	502	116,648	Palm Beach	152,951	269,732	3411	433,186
Flagler	12,613	13,897	83	27,111	Pasco	68,582	69,564	570	142,731
Franklin	2454	2046	33	4644	Pinellas	184,825	200,630	1013	398,472
Gadsden	4767	9735	38	14,727	Polk	90,295	75,200	533	168,607
Gilchrist	3300	1910	29	5395	Putnam	13,447	12,102	148	26,222
Glades	1841	1442	9	3365	Santa Rosa	36,274	12,802	311	50,319
Gulf	3550	2397	71	6144	Sarasota	83,100	72,853	305	160,942
Hamilton	2146	1722	23	3964	Seminole	75,677	59,174	194	137,634
Hardee	3765	2339	30	6233	St. Johns	39,546	19,502	229	60,746
Hendry	4747	3240	22	8139	St. Lucie	34,705	41,559	124	77,989
Hernando	30,646	32,644	242	65,219	Sumter	12,127	9637	114	22,261
Highlands	20,206	14,167	127	35,149	Suwannee	8006	4075	108	12,457
Hillsborough	180,760	169,557	847	360,295	Taylor	4056	2649	27	6808
Holmes	5011	2177	76	7395	Union	2332	1407	37	3826
Indian River	28,635	19,768	105	49,622	Volusia	82,357	97,304	498	183,653
Jackson	9138	6868	102	16,300	Wakulla	4512	3838	46	8587
Jefferson	2478	3041	29	5643	Walton	12,182	5642	120	18,318
Lafayette	1670	789	10	2505	Washington	4994	2798	88	8025
Lake	50,010	36,571	289	88,611	Absentee	1575	836	5	2490

Chapter 12

SIMULATION

12.1 What Is Simulation?

Simulation is a way to use high-speed computer power to substitute for analytical calculation. The law of large numbers tells us that if we observe a large sample of i.i.d. random variables with finite mean, then the average of these random variables should be close to their mean. If we can get a computer to produce such a large i.i.d. sample, then we can average the random variables instead of trying (and possibly failing) to calculate their mean analytically. For a specific problem, one needs to figure out what types of random variables one needs, how to make a computer produce them, and how many one needs in order to have any confidence in the numerical result. Each of these issues will be addressed to some extent in this chapter.

Proof of Concept

We begin with some simple examples of simulation to answer questions that we can already answer analytically just to show that simulation does what it advertises. Also, these simple examples will raise some of the issues to which one must attend when trying to answer more difficult questions using simulation.

Example 12.1.1 The Mean of a Distribution. The mean of the uniform distribution on the interval [0, 1] is known to be 1/2. If we had available a large number of i.i.d. uniform random variables on the interval [0, 1], say, $X_1, \ldots, X_n$, the law of large numbers says that $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ should be close to the mean 1/2. Table 12.1 gives the averages of several different simulated samples of size n from the uniform distribution on [0, 1] for several different values of n. It is not difficult to see that the averages are close to 0.5 in most cases, but there is quite a bit of variation, especially for $n = 100$. There seems to be less variation for $n = 1000$, and even less for the two largest values of n. ◄

Example 12.1.2 A Normal Probability. The probability that a standard normal random variable is at least 1.0 is known to be 0.1587. If we had available a large number of i.i.d. standard normal random variables, say, $X_1, \ldots, X_n$, we could create Bernoulli random variables $Y_1, \ldots, Y_n$ defined by $Y_i = 1$ if $X_i \geq 1.0$ and $Y_i = 0$ if not. Then the law of large numbers says that $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ should be close to the mean of Y_i, namely,

Table 12.1 Results of several different simulations in Example 12.1.1

n	Replications of the simulation				
100	0.485	0.481	0.484	0.569	0.441
1000	0.497	0.506	0.480	0.498	0.499
10,000	0.502	0.501	0.499	0.498	0.498
100,000	0.502	0.499	0.500	0.498	0.499

Table 12.2 Results of several different simulations in Example 12.1.2

n	Replications of the simulation				
100	0.16	0.18	0.17	0.22	0.14
1000	0.135	0.171	0.174	0.159	0.171
10,000	0.160	0.163	0.158	0.152	0.156
100,000	0.158	0.158	0.158	0.159	0.161

$\Pr(X_i \geq 1.0) = 0.1587$. Notice that $\overline{Y}$ is merely the proportion of the simulated X_i values that are at least 1.0. Table 12.2 gives the proportions of $X_i \geq 1.0$ for several different simulated samples of size n from the standard normal distribution for several different values of n. It is not difficult to see that the proportions are somewhat close to 0.1587, but there is still quite a bit of variability from one simulation to the next. ◀

As we mentioned earlier, there is no need for simulation in the above examples. These were just to illustrate that simulation can do what it claims. However, one needs to be aware that, no matter how large a sample is simulated, the average of an i.i.d. sample of random variables is not necessarily going to be equal to its mean. One needs to be able to take the variability into account. The variability is apparent in Tables 12.1 and 12.2. We shall address the issue of the variability of simulations later in the chapter.

The reader might also be wondering how we obtained all of the uniform and normal random variables used in the examples. Virtually every commercial statistical software package has a simulator for i.i.d. uniform random variables on the interval [0, 1]. Later in the chapter, we shall discuss ways to make use of these for simulating other distributions. One such method was already discussed in Chapter 3 on page 170.

Examples in which Simulation Might Help

Next, we present some examples where the basic questions are relatively simple to describe, but analytic solution would be tedious at best.

Example 12.1.3 Waiting for a Break. Two servers, A and B, in a fast-food restaurant start serving customers at the same time. They agree to meet for a break after each of them has

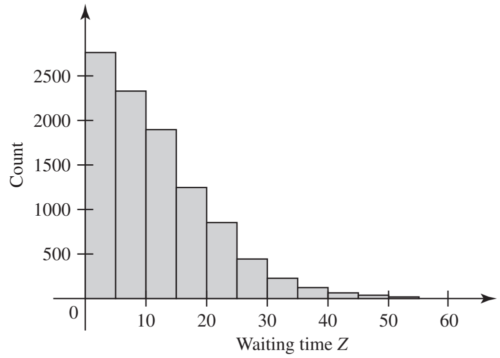


Figure 12.1 Histogram of sample of 10,000 simulated waiting times Z in Example 12.1.3.

served 10 customers. Presumably, one of them will finish before the other and have to wait. How long, on average, will one of the servers have to wait for the other?

Suppose that we model all service times, regardless of the server, as i.i.d. random variables having the exponential distribution with parameter 0.3 customers per minute. Then the time it takes one server to server 10 customers has the gamma distribution with parameters 10 and 0.3. Let X be the time it takes A to serve 10 customers, and let Y be the time it takes B to serve 10 customers. We are asked to compute the mean of $|X - Y|$. The most straightforward way of finding this mean analytically would require a two-dimensional integral over the union of two non-rectangular regions.

On the other hand, suppose that a computer can provide us with as many independent gamma random variables as we desire. We can then obtain a pair (X, Y) and compute $Z = |X - Y|$. We then repeat this process independently as many times as we want and average all the observed Z values. The average should be close to the mean of Z.

Without going into details, we actually simulated 10,000 pairs of (X, Y) values and averaged the resulting Z values to get 11.71 minutes. A histogram of the simulated Z values is in Fig. 12.1. As a confidence builder, we simulated another 10,000 pairs and got an average of 11.77. ◀

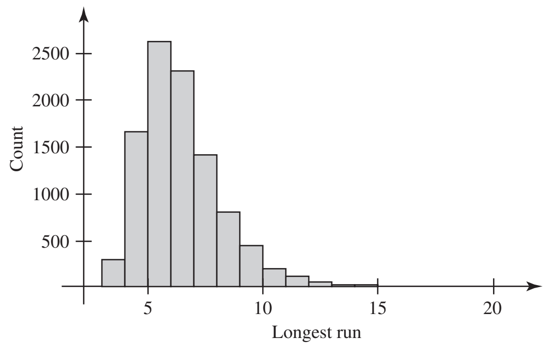
Example 12.1.4

Long Run of Heads. You overheard someone say that they just got 12 consecutive heads while flipping a seemingly fair coin. The probability of getting 12 heads in a row in 12 independent flips of a fair coin is $(0.5)^{12}$, a very small number. If the person had obtained 12 tails in a row, you probably would have heard about that instead. Even so, the probability of 12 of the same side is only $(0.5)^{11}$. But then you learn that the person actually flipped the coin 100 times, and the 12 heads in a row appeared somewhere during those 100 flips. Presumably, you are less surprised to learn that the person got a run of 12 of the same side somewhere in a sequence of 100 flips. But how much larger is the probability of a run of 12 when one flips 100 times?

Suppose that we can make a computer flip a fair coin as many times as we wish. We could ask it to flip 100 times and then check whether there was a run of length 12 or more. Let $X = 1$ if there is a run of 12 or more, and let $X = 0$ if not. We then repeat this process independently as many times as we want and average all the observed X values. The average should be close to the mean of X, which is the probability of obtaining a run of 12 or more in 100 flips.

Without going into details, Fig. 12.2 shows a histogram of the longest runs in 10,000 repetitions of the experiment described above. For each of the 10,000 runs, we calculated X as above and found the average to be 0.0214, still a small number,

Figure 12.2 Histogram of sample of 10,000 longest runs (head or tail). Each run was observed in 100 flips of a fair coin.



but not nearly so small as $(0.5)^{11}$. We also repeated the calculation of the average with another 10,000 sets of 100 flips and got 0.0229. ◀

A number of details were left out of exactly how the simulations were performed in the above examples. However, it is clear what random variables we wanted to observe, namely, Z in Example 12.1.3 and X in Example 12.1.4. Many simulations can address more than one question. For instance, in Example 12.1.4, we recorded the 10,000 lengths of the longest runs even though our primary interest was in whether or not the longest run was 12 or more. We could also have tried to calculate the expected length of the longest run or other properties of the distribution of the longest run. In Example 12.1.3, we could have tried to approximate the probability that one person has to wait at least 15 minutes, etc.

Figures 12.1 and 12.2 illustrate that there is variation among the 10,000 repetitions of a simulated experiment. Furthermore, each of the examples showed that a complete rerunning of all 10,000 simulated experiments can be expected to produce a different answer to each of our questions. How much different the answers should be is a matter that we shall address in Sec. 12.2, where we use the Chebyshev inequality and the central limit theorem to help us decide how many times to repeat the basic experiment. Exactly how one simulates 100 flips of a coin or a pair of gamma random variables will be taken up in Sec. 12.3.

Summary

Suppose that we want to know the mean of some function g of a random variable or random vector W. For instance, in Example 12.1.3 we can let $W = (X, Y)$ and $g(W) = |X - Y|$. If a computer can supply a large number of i.i.d. random variables (or random vectors) with the distribution of W, one can use the average of the simulated values of $g(W)$ to approximate the mean of $g(X)$. One must be careful to take the variability in $g(W)$ into account when deciding how much confidence to place in the approximation.

Exercises

For each of the exercises in this section, you could also perform the simulations described with various numbers of replications if you have appropriate computer software available. Most of the distributions involved are commonly available in computer software. If a distribution is not available, the simulations can wait until methods for simulating specific distributions are introduced in Sec. 12.3.

1. Assume that one can simulate as many i.i.d. exponential random variables with parameter 1 as one wishes. Explain how one could use simulation to approximate the mean of the exponential distribution with parameter 1.

2. If X has the p.d.f. $1/x^2$ for $x > 1$, the mean of X is infinite. What would you expect to happen if you simulated a large number of random variables with this p.d.f. and computed their average?

3. If X has the Cauchy distribution, the mean of X does not exist. What would you expect to happen if you simulated a large number of Cauchy random variables and computed their average?

4. Suppose that one can simulate as many i.i.d. Bernoulli random variarbles with parameter p as one wishes. Explain how to use these to approximate the mean of the geometric distribution with parameter p.

5. Two servers A and B in a fast-food restaurant each start their first customers at the same time. After finishing her second customer, A notices that B has not yet finished his first customer. A then chides B for being slow, and B responds that A just got a couple of easier customers. Suppose that we model all service times, regardless of the server, as i.i.d. random variables having the exponential distribution with parameter 0.4. Let X be the sum of the first two service times for server A, and let Y be the first service time for server B. Assume that you can simulate as many i.i.d. exponential random variables with parameter 0.4 as you wish.

a. Explain how to use such random variables to approximate $\Pr(X < Y)$.

b. Explain why $\Pr(X < Y)$ is the same no matter what the common parameter is of the exponential distribuions. That is, we don't need to simulate exponentials with parameter 0.4. We could use any parameter that is convenient, and we should get the same answer.

c. Find the joint p.d.f. of X and Y, and write the two-dimensional integral whose value would be $\Pr(X < Y)$.

12.2 Why Is Simulation Useful?

Statistical simulations are used to estimate features of distributions such as means of functions, quantiles, and other features that we cannot compute in closed form. When using a simulation estimator, it is good to compute a measure of how precise the estimator is, in addition to the estimate itself.

Examples of Simulation

Simulation is a technique that can be used to help shed light on how a complicated system works even if detailed analysis is unavailable. For example, engineers can simulate traffic patterns in the vicinity of a construction project to see what effects various proposed restrictions might have. A physicist can simulate the behavior of gas molecules under conditions that are covered by no known theory. Statistical simulations are used to estimate probabilistic features of our models that we cannot compute analytically. Because simulation introduces an element of randomness into an analysis, it is sometimes called *Monte Carlo analysis*, named after the famous European gambling center.

Example 12.2.1

The M.S.E. of the Sample Median. Suppose that we are about to observe a random sample of size n from a Cauchy distribution centered at an unknown value μ. The p.d.f. of each observation is

$$f(x) = \frac{1}{\pi}(1 + [x - \mu]^2)^{-1},$$

and the parameter μ is the median of the distribution. Suppose that we are interested in how well the sample median M performs as an estimator of μ. In particular, we want to calculate the M.S.E. $E([M - \mu]^2)$. If we could generate a sample of n random variables from a Cauchy distribution centered at μ, we could compute the sample median M and calculate $Y = (M - \mu)^2$. The M.S.E. is then $\theta = E(Y)$. If we could

generate a large number v of i.i.d. random variables with the same distribution as Y, say, $Y^{(1)}, \ldots, Y^{(v)}$, then the law of large numbers would tell us that $Z = \frac{1}{v}\sum_{i=1}^{v} Y^{(i)}$ should be close to θ. To do this, we could generate nv i.i.d. Cauchy random variables centered at μ. Then we could divide them into v sets of n each and use each set of n to compute a sample median $M^{(i)}$ for $i = 1, \ldots, v$ and then compute $Y^{(i)} = (M^{(i)} - \mu)^2$. This is actually how several of the numbers in the tables in Sec. 10.7 were computed. These tables contain the M.S.E.'s of various estimators computed from random samples with various distributions. For example, the numbers corresponding to the sample median in Table 10.39 on page 675 are precisely what we have been discussing in this example. ◄

Note: Notation to Distinguish Simulations. We shall use superscripts in parentheses to distinguish different simulated values of the same random variable from each other. For instance, in Example 12.2.1, we used $Y^{(i)}$ to stand for the ith simulated value of Y. In what follows, we may be simulating subscripted random variables. For example, $\mu_i^{(j)}$ would stand for the jth simulated value of μ_i.

Example 12.2.1 illustrates the main features of many statistical simulations. Suppose that the quantity in which we are interested can be expressed as the expected value of some random variable that has the distribution F. Then we should try to generate a large sample of random variables with the distribution F and average them. It is often the case, as in Example 12.2.1, that the distribution F is itself very complicated. In such cases, we need to construct random variables with the distribution F from simpler random variables whose distributions are more familiar. In Example 12.2.1, the M.S.E. is the mean of the random variable $Y = (M - \mu)^2$, where M is itself the sample median of a sample of n Cauchy random variables centered at μ. We cannot easily simulate a random variable with the distribution of Y in one step, but we can simulate n Cauchy random variables and then find their sample median M and finally compute $Y = (M - \mu)^2$, which will have the desired distribution. We then repeat the simulation of Y many times.

Not all statistical simulations involve the mean of a random variable.

Example 12.2.2 The Median of a Complicated Distribution. Let X be an exponential random variable with unknown parameter μ. Suppose that μ has a distribution with the p.d.f. g. We are interested in the median of X. The marginal distribution of X has the p.d.f.

$$f(x) = \int_0^\infty \mu e^{-\mu x} g(\mu) d\mu.$$

This integral might not be one that we can compute. However, suppose that we can generate a large sample of random variables $\mu^{(1)}, \ldots, \mu^{(v)}$ having the p.d.f. g. Then, for each $i = 1, \ldots, v$, we can simulate $X^{(i)}$ having the exponential distribution with parameter $\mu^{(i)}$. The random variables $X^{(1)}, \ldots, X^{(v)}$ would then be a random sample from the distribution with the p.d.f. f. The median of the sample $X^{(1)}, \ldots, X^{(v)}$ should be close to the median of the distribution with the p.d.f. f. ◄

Example 12.2.3 A Clinical Trial. Consider the four treatment groups described in Example 2.1.4 on page 57. For $i = 1, 2, 3, 4$, let P_i be the probability that a patient in treatment group i will not relapse after treatment. We might be interested in how likely it is that the P_i's differ by certain amounts. We might assume that the P_i's are independent a priori with beta distributions having parameters α_0 and β_0. The posterior distributions of the P_i's are also independent beta distributions with parameters $\alpha_0 + x_i$ and $\beta_0 + n_i - x_i$, where n_i is the number of subjects in group i, and x_i is the number of patients in group

i who do not relapse. We could simulate a large number v of vectors (P_1, P_2, P_3, P_4) with the above beta distributions. Then we could try to answer any question we wanted about the posterior distribution of (P_1, P_2, P_3, P_4). For example, we could estimate $\Pr(P_i > P_4)$ for $i = 1, 2, 3$, where $i = 4$ stands for the placebo group. This probability tells us how likely it is that each treatment is better than no treatment. We could estimate $\Pr(P_i > P_4)$ by finding the proportion of sampled (P_1, P_2, P_3, P_4) vectors in which the ith coordinate is greater than the fourth coordinate. We could also estimate the probability that P_i is the largest, or the probability that all four P_i are within ϵ of each other. ◀

Example 12.2.4 Comparing Two Normal Means with Unequal Variances. On page 593 in Chapter 9, we considered how to test hypotheses concerning the means of two different normal distributions when the variances are unknown and different. This problem has a relatively simple solution in the Bayesian framework using simulation. Our parameters will be μ_x, τ_x, μ_y, and τ_y. Conditional on the parameters, let $X_1, \ldots, X_m$ be i.i.d. having the normal distribution with mean μ_x and precision τ_x. Also let $Y_1, \ldots, Y_n$ be i.i.d. (and independent of the X's) having the normal distribution with mean μ_y and precision τ_y. Assume that we use natural conjugate priors for the parameters with (μ_x, τ_x) independent of (μ_y, τ_y) in the prior distribution. (It is not necessary for the X parameters to be independent of the Y parameters, but it makes the presentation simpler.) Sec. 8.6 contains details on how to obtain the posterior distributions of the parameters. Since the X data and X parameters are independent of the Y data and Y parameters, we can calculate each posterior distribution separately. Let the hyperparameters of the posterior distribution of (μ_x, τ_x) be α_{x1}, β_{x1}, μ_{x1}, and λ_{x1}. Similarly, let the hyperparameters of the posterior distribution of (μ_y, τ_y) be α_{y1}, β_{y1}, μ_{y1}, and λ_{y1}. In order to test hypotheses about $\mu_x - \mu_y$, we need the posterior distribution of $\mu_x - \mu_y$. This distribution is not analytically tractable. If we can simulate a large collection of parameter vectors from their joint posterior distribution, we can compute $\mu_x - \mu_y$ for each sampled vector, and these values will form a sample from the posterior distribution of $\mu_x - \mu_y$. To be more specific, let v be a large number, and for each $i = 1, \ldots, v$, we want to simulate $(\mu_x^{(i)}, \mu_y^{(i)}, \tau_x^{(i)}, \tau_y^{(i)})$ from the joint posterior distribution. To do this, we need to simulate independent gamma random variables $\tau_x^{(i)}$ and $\tau_y^{(i)}$ with the appropriate posterior distributions. After simulating these, we can simulate $\mu_x^{(i)}$ from the normal distribution with mean μ_{x1} and variance $1/(\lambda_{x1}\tau_x^{(i)})$. Similarly, we can simulate $\mu_y^{(i)}$ from the normal distribution with mean μ_{y1} and variance $1/(\lambda_{y1}\tau_y^{(i)})$. Then $\mu_x^{(i)} - \mu_y^{(i)}$ for $i = 1, \ldots, v$ is a sample from the posterior distribution of $\mu_x - \mu_y$. We shall illustrate this methodology in Example 12.3.8 after we discuss some methods for simulating pseudo-random numbers with various distributions. ◀

The simulation in Example 12.2.4 can be extended in a straightforward fashion to a comparison of three or more normal distributions with unequal variances. With more than two means to compare, questions arise about what exactly to calculate to summarize the comparison. That is, there is not just one difference like $\mu_x - \mu_y$ that captures the differences between three or more means. We shall consider this situation in more detail in Examples 12.3.7 and 12.5.6.

Example 12.2.5 Estimating a Standard Deviation. Let X be a random variable whose standard deviation θ is important to estimate. Suppose that we cannot calculate θ in closed form, but we can simulate many pseudo-random values $X^{(1)}, \ldots, X^{(v)}$ with the same distribution

as X. Then we could compute the sample standard deviation

$$S_v = \left(\frac{1}{v} \sum_{i=1}^{v} (X^{(i)} - \overline{X})^2 \right)^{1/2},$$

as an estimator of θ, where $\overline{X} = \frac{1}{v} \sum_{i=1}^{v} X^{(i)}$. Since S_v is not an average, the law of large numbers does not tell us that it converges in probability to θ. However, if we let $Y^{(i)} = X^{(i)2}$, we can rewrite S_v as $(\overline{Y} - \overline{X}^2)^{1/2}$. In this form, we see that $S_v = g(\overline{X}, \overline{Y})$, where $g(x, y) = (y - x^2)^{1/2}$. Notice that g is continuous at every point (x, y) such that $y \ge x^2$. The law of large numbers tells us that $\overline{Y}$ converges in probability to $E(X^2)$ and that $\overline{X}$ converges in probability to $E(X)$. Since $E(X^2) \ge E(X)$, we can apply Exercise 16 in Sec. 6.2 to conclude that S_v converges in probability to $g(E(X), E(X^2)) = \theta$. ◂

All of the examples above involve the generation of a large number of random variables with specific distributions. Some discussion of this topic appeared in Chapter 3 beginning on page 170. Sections 12.3 and 12.5 will also discuss methods for generating random variables with specific distributions. Sections 12.4 and 12.6 will present particular classes of problems in which statistical simulation is used successfully.

Which Mean Do You Mean?

Simulation analyses add an additional layer of probability distributions and sampling distributions of statistics to an already probability-laden statistical analysis. A typical statistical analysis involves a probability model for a random sample of data $X_1, \dots, X_n$. This probability model specifies the distribution of each X_i, and this distribution might have parameters such as its mean, median, variance, and other measures that we are interested in estimating or testing. We then form statistics (functions of the data), say, $\boldsymbol{Y}$. These functions might include sample versions of the very parameters that we wish to estimate, such as a sample mean, sample median, sample variance, and the like. The distribution of $\boldsymbol{Y}$ has been called its sampling distribution. This sampling distribution also might have a mean, median, variance, and other measures that we need to calculate or deal with in some way. So far, we have three versions of mean, median, variance, and others, and we have not even begun discussing simulation.

A simulation analysis might be used to try to estimate a parameter θ of the sampling distribution of the statistics $\boldsymbol{Y}$. Typically, one would simulate i.i.d. pseudo-random $\boldsymbol{Y}^{(1)}, \dots, \boldsymbol{Y}^{(v)}$ each with the same distribution as (the sampling distribution of) $\boldsymbol{Y}$. We then compute a summary statistic Z of $\boldsymbol{Y}^{(1)}, \dots, \boldsymbol{Y}^{(v)}$ and use Z to estimate θ. This Z might itself be a sample mean, sample median, sample variance, or other measure of the $\boldsymbol{Y}^{(1)}, \dots, \boldsymbol{Y}^{(v)}$ sample. The distribution of Z will be called its *simulation distribution* or *Monte Carlo distribution*. Features of the simulation distribution, such as its mean, median, and variance, will be called the simulation mean, simulation median, and simulation variance to make clear to which level we have climbed in this ever-expanding tree of terminology. Here is an example to illustrate all of the various levels.

Example 12.2.6 Five or More Variances. Let $X_1, \dots, X_n$ be i.i.d. random variables with a continuous distribution having c.d.f. F. Let ψ denote the variance of X_i. Suppose that we decide to use the sample variance $Y = \sum_{i=1}^{n} (X_i - \overline{X})^2/n$ to estimate ψ. As part of deciding

Table 12.3 Levels of probability distributions, statistics, and parameters in a typical simulation analysis

Distribution (D) or sample (S)	Parameter (P) or statistic (S)
(D) Population distribution F	(P) Mean, variance, median, etc. ψ
(S) Sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from F	(S) Estimator Y of ψ based on $\boldsymbol{X}$, e.g., sample mean, sample variance, sample median, etc.
(D) Sampling distribution G of Y	(P) Mean, variance, median, etc., θ of the sampling distribution of Y
(S) Simulated sample $\boldsymbol{Y} = (Y^{(1)}, \dots, Y^{(v)})$ from G	(S) Estimator Z of θ based on $\boldsymbol{Y}$, e.g., sample mean, sample variance, sample median, etc., of $\boldsymbol{Y}$.
(D) Simulation distribution H of Z	(P) Variance of simulation distribution (simulation variance)
(S) Simulated data (differs by example)	(S) Estimator of simulation variance, (depends on specific example)

how good Y is as an estimator of ψ, we are interested in its variance $\theta = \text{Var}(Y)$. That is, θ is the variance of the sampling distribution of Y. Suppose that we cannot calculate θ in closed form, but suppose that it is easy to simulate from the distribution F. We might then simulate nv values $X_i^{(j)}$ for $j = 1, \dots, v$, $i = 1, \dots, n$. For each j, we compute the sample variance $Y^{(j)}$ of the sample $X_1^{(j)}, \dots, X_n^{(j)}$. That is, $Y^{(j)} = \sum_{i=1}^{n}(X_i^{(j)} - \overline{X}^{(j)})^2/n$. The $Y^{(j)}$ values all have the same distribution as Y itself, the sampling distribution of Y. Since we are interested in Var(Y), we might compute the sample variance Z of the sample $Y^{(1)}, \dots, Y^{(v)}$. That is, $Z = \sum_{i=1}^{v}(Y^{(i)} - \overline{Y})^2/v$. We would then use Z to estimate θ. If Z is large, it suggests that Y has large variance, and so Y is not a very good estimator of ψ. Unless we are willing to collect more data or search for a better estimator, we are stuck with a poor estimator of ψ.

Finally, Z might not be a good estimator of θ because our simulation size v might not be large enough. If this is the case, we can simulate more $Y^{(j)}$ values. That is, we can increase the simulation size v to get a better simulation estimator of θ. (This will not make Y a better estimator of ψ, but it will give us a better idea of how good or bad an estimator it is.) Hence, we shall also try to estimate the variance of Z (its simulation variance). Precisely how to do this varies from one example to the next, so we shall not give any details here. However, we shall explain how to estimate the simulation variance of Z for the most popular types of simulation later in this section.

This estimation of variance has to end somewhere, and we shall end it with Var(Z). That is, we shall *not* try to assess how good our estimator of Var(Z) is. All of these levels of distributions and estimation are illustrated in Table 12.3. ◄

Example 12.2.6 is not intended to illustrate any simulation methodology. It is intended to illustrate the various levels at which probability concepts (such as variance) and their sample versions enter into a simulation study of a statistical analysis. It is important to be able to tell which variance or which sample variance is being discussed if one is to avoid becoming hopelessly confused. In this chapter, we shall focus on the features of the simulated samples, in particular the simulation distribution of statistics computed from the simulated samples. However, our examples will necessarily involve parameters and statistics that arose at earlier levels. Furthermore, the analysis of a simulation distribution will make use of the same methods (central

limit theorem, law of large numbers, delta method, etc.) that we learned how to use with nonsimulated data.

Assessing Uncertainty about Simulation Results

The last step in Example 12.2.6 (summarized in the last two rows of Table 12.3) is an important part of every simulation analysis. That is, we should always attempt to assess the uncertainty in a simulation. This uncertainty is most easily assessed via the simulation variance of the simulated quantity. For instance, in Example 12.2.1, let $v = 1000$ and $\theta = 0$. We can create 1000 samples of n Cauchy random variables, calculate $M^{(i)}$, the median of the ith sample, and compute the value $Y^{(i)} = (M^{(i)} - 0)^2$. We can then average the 1000 values of $Y^{(i)}$. We could repeat this exercise several times, and we would not get the same result every time. This is due to the fact that, even with a large v like 1000, an estimator such as $Z = \frac{1}{v}\sum_{i=1}^{v} Y^{(i)}$ is still a random variable with positive variance (its simulation variance). The smaller the simulation variance is, the more certain we can be that our estimator Z is close to what we are trying to estimate. But we need to estimate or bound the simulation variance before we can assess the amount of uncertainty. How we estimate the simulation variance of a result Z depends on whether Z is an average of simulated values, a smooth function of one or more averages, or a sample quantile of simulated values. The square root of our estimate of the simulation variance will be called the *simulation standard error*, and it is an estimate of the simulation standard deviation of Z. The simulation standard error is a popular way to summarize uncertainty about a simulation for two reasons. First, it has the same units of measurement as the quantity that was estimated (unlike the simulation variance). Second, the simulation standard error is useful for saying how likely it is that the simulation estimator is close to the parameter being estimated. We shall explain this second point in more detail after we show how to calculate the simulation standard error in several common cases.

Example 12.2.7 **The Simulation Standard Error of an Average.** Suppose that the goal of the simulation analysis is to estimate the mean θ of some random variable Y. The simulation estimator Z will generally be the average of a large number of simulated values. A straightforward way to estimate the simulation variance for an average is the following: Suppose that we simulate some quantity Y a large number v of times in order to estimate the mean θ. That is, suppose that we simulate independent $Y^{(1)}, \dots, Y^{(v)}$ for large v. Suppose also that the estimator of θ is $Z = \frac{1}{v}\sum_{i=1}^{v} Y^{(i)}$, and each $Y^{(i)}$ has mean θ and finite variance σ^2. The sample standard deviation of the sample $Y^{(1)}, \dots, Y^{(v)}$ is the square root of the sample variance, namely,

$$\hat{\sigma} = \left(\frac{1}{v} \sum_{i=1}^{v} (Y^{(i)} - \overline{Y})^2 \right)^{1/2}. \tag{12.2.1}$$

If v is large, then $\hat{\sigma}$ should be close to σ. The central limit theorem says that Z should have approximately the normal distribution with mean θ and variance σ^2/v. Since we usually do not know σ^2, we shall estimate it by $\hat{\sigma}^2$. This makes our estimator of the simulation variance of Z equal to $\hat{\sigma}^2/v$, and the simulation standard error is $\hat{\sigma}/v^{1/2}$. ◀

Example 12.2.8 **The Simulation Standard Error of a Smooth Function of Another Estimator.** Sometimes, after estimating a quantity ψ, we also wish to estimate a smooth function of it: $g(\psi)$. For example, we might need to estimate the square root or the logarithm of some

mean. Or, we might have estimated a variance θ^2, and now we want an estimator of θ, the corresponding standard deviation. In general, suppose that the parameter that we wish to estimate by simulation is $\theta = g(\psi)$, where we already have an estimator W of ψ. Suppose further that our estimator W has approximately the normal distribution with mean ψ and variance σ^2/v, where v is large compared to σ^2. Finally, suppose that we also have an estimator $\hat{\sigma}$ of σ that we obtained while calculating W. For example, W might itself be the average of v i.i.d. simulated random variables $Y^{(i)}$ with mean ψ and variance σ^2. In this case, Eq. (12.2.1) will be our estimator of σ. Let $Z = g(W)$ be our estimator of θ. The delta method (see Sec. 6.3) says that Z has approximately the normal distribution with mean $\theta = g(\psi)$ and variance $[g'(\psi)]^2\sigma^2/v$. For example, if $g(\psi) = \psi^{1/2}$, then $W^{1/2}$ has approximately the normal distribution with mean $\psi^{1/2}$ and variance $\sigma^2/[4\psi v]$. We already have estimates of σ and ψ, so our simulation standard error of Z is $|g'(W)|\hat{\sigma}/v^{1/2}$. ◀

Example 12.2.9 **The Simulation Standard Error of a Sample Quantile.** Suppose that the goal of a simulation analysis is to estimate the p quantile θ_p of some distribution G. Typically, we simulate a large number v of pseudo-random values $Y^{(1)}, \dots, Y^{(v)}$ with distribution G and use the sample p quantile as our estimator. On page 676, we pointed out that the sample p quantile from a large random sample of size m has approximately the normal distribution with mean θ_p and variance $p(1-p)/[mg^2(\theta_p)]$, where g is the p.d.f. of the distribution G. All we care about right now is that this approximate variance has the form σ^2/m, where $\sigma^2 = p(1-p)/g^2(\theta_p)$ is some number that does not depend on m. Suppose that we simulate k independent random samples each of size m from the distribution G. Typically, this is done by choosing the size v of the original simulated sample $Y^{(1)}, \dots, Y^{(v)}$ to be $v = km$, and then splitting the v simulated values into k subsamples of size m each. Compute the sample p quantile of each of the k random samples and call these simulated sample p quantiles $Z_1, \dots, Z_k$. To make use of the approximate normal distribution for the sample quantiles, m needs to be large. Next, compute the sample standard deviation of $Z_1, \dots, Z_k$:

$$S = \left(\frac{1}{k} \sum_{i=1}^{k} (Z_i - \overline{Z})^2 \right)^{1/2}, \qquad (12.2.2)$$

where $\overline{Z}$ is the average of the k sample p quantiles. If we treat each Z_i as a single simulation, then S^2 is an estimator of the variance of Z_i. But we just pointed out that the variance of Z_i is approximately σ^2/m. Hence, S^2 is an estimator of σ^2/m. In other words, an estimator of σ is $\hat{\sigma} = m^{1/2}S$. Finally, combine all k samples into a single sample of size $v = km$, and compute the sample p quantile Z as our Monte Carlo estimator of θ_p. As we noted earlier, Z has approximately the normal distribution with mean θ_p and variance σ^2/v. We just constructed an estimator $\hat{\sigma}$ of σ, so our estimator of the simulation variance of Z is $\hat{\sigma}^2/v = mS^2/v = S^2/k$, and the simulation standard error is $S/k^{1/2}$. ◀

Example 12.2.10 **The Simulation Standard Error of a Sample Variance.** Suppose that the goal of a simulation analysis is to estimate the variance θ of some estimator Y. (Example 12.2.6 was based on such a situation.) Suppose that we simulate $Y^{(1)}, \dots, Y^{(v)}$ and use $Z = \frac{1}{v}\sum_{i=1}^{v}(Y^{(i)} - \overline{Y})^2$ to estimate θ. We now need to estimate the simulation variance of Z. We shall rewrite Z as a smooth function of two averages and then apply a two-dimensional generalization of the delta method (see Exercise 12) in order to estimate the simulation variance. Let $W^{(i)} = Y^{(i)2}$ so that $Z = \overline{W} - \overline{Y}^2$, where $\overline{W}$ is

the average of $W^{(1)}, \ldots, W^{(v)}$. Now Z is a smooth function of two averages. The two-dimensional delta method developed in Exercise 12 can be applied. (The details for this very case can be derived in Exercise 13.) The results of Exercise 13 provide the following approximation to the asymptotic variance of Z. First, compute the sample variance of $W^{(1)}, \ldots, W^{(v)}$ and call it V. Next, compute the sample covariance between the Y's and W's:

$$C = \frac{1}{v}\sum_{i=1}^{v}(Y^{(i)} - \overline{Y})(W^{(i)} - \overline{W}).$$

The estimator of $\text{Var}(Z)$ is then

$$\widehat{\text{Var}(Z)} = \frac{1}{v}\left(4\overline{Y}^2 Z - 4\overline{Y}C + V\right). \tag{12.2.3}$$

Also, the simulation distribution of Z is approximately the normal distribution with mean θ and variance that is estimated by Eq. (12.2.3). The simulation standard error is the square root of (12.2.3). ◄

Do We Have Enough Simulations? Let Z be our Monte Carlo estimator of some parameter θ based on v simulations. Now that we are able to estimate the simulation variance of Z, we can begin to answer questions about how close we think Z is to θ. We can also try to see if we need to do more simulations in order to be confident that Z is close enough to θ. Suppose, as in all of the cases considered so far, that Z has approximately the normal distribution with mean θ and variance σ^2/v, where σ^2 is a number that does not depend on the simulation size. For each $\epsilon > 0$,

$$\Pr(|Z - \theta| \le \epsilon) \approx 2\Phi\left(\epsilon v^{1/2}/\sigma\right) - 1, \tag{12.2.4}$$

where Φ is the standard normal c.d.f. We can use this type of approximation to help us to say how likely it is that Z is close to θ. We can replace $v^{1/2}/\sigma$ by 1 over the simulation standard error of Z in Eq. (12.2.4) to approximate the probability that $|Z - \theta| \le \epsilon$. We can also use (12.2.4) to decide how many more simulations to do if v was not large enough. For example, suppose that we want the probability in Eq. (12.2.4) to be γ. Then we should let

$$v = \left[\Phi^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma}{\epsilon}\right]^2. \tag{12.2.5}$$

Since we will hardly ever know σ ahead of time, it is common to estimate it by doing a preliminary simulation of size v_0 and computing $\hat{\sigma}$ based on that preliminary simulation.

Example 12.2.11 The M.S.E. of the Sample Median. It is not difficult to see that we can take $\mu = 0$ in Example 12.2.1 without loss of generality. The reason is the following: Let $M^{(i)}$ be the sample median of $X_1^{(i)}, \ldots, X_n^{(i)}$ where each $X_j^{(i)}$ is a Cauchy random variable centered at μ. Then $M^{(i)} - \mu$ is also the sample median of $X_1^{(i)} - \mu, \ldots, X_n^{(i)} - \mu$, and each $X_j^{(i)} - \mu$ is a Cauchy random variable centered at 0. Because our calculation is based on the values $Y^{(i)} = (M^{(i)} - \mu)^2$ for $i = 1, \ldots, v$, we get the same result whether $\mu = 0$ or not. So, let $\mu = 0$. This makes $Y^{(i)} = M^{(i)2}$, and σ^2 is now the variance of $M^{(i)2}$. (Even though a Cauchy random variable does not even have a first moment defined, it can be shown that the sample median of at least nine i.i.d. Cauchy random variables has a finite fourth moment.) Suppose that we want our estimator $Z = \overline{Y}$ of θ to be within $\epsilon = 0.01$ of θ with probability $\gamma = 0.95$. That

is, we want $\Pr(|Z-\theta|\le 0.01)=0.95$. Since Z is an average, we can compute an estimate $\hat{\sigma}$ of σ by using Eq. (12.2.1). Suppose that we simulate $v_0=1000$ samples of size $n=20$ from a Cauchy distribution, compute the 1000 values of $Y^{(i)}$, and then compute $\hat{\sigma}=0.3892$. According to Eq. (12.2.5) with σ replaced by 0.3892, we need $v=[1.96\times 0.3892/0.01]^2=5820$. Hence, we need approximately 4820 additional simulations. ◀

After performing any additional simulations suggested by Eq. (12.2.5), one should recompute $\hat{\sigma}$. If it is much larger than the preliminary estimate, then additional simulations should be performed.

Example 12.2.12 The Median of a Complicated Distribution. In Example 12.2.2, suppose that the p.d.f. g is the p.d.f. of the gamma distribution with parameters 3 and 1. Suppose that we want the probability to be 0.99 that our estimator of the median is within 0.001 of the true median. We begin with an initial simulation of size v_0 =10,000. We then simulate $\mu^{(1)},\dots,\mu^{(10,000)}$ from the gamma distribution with parameters 3 and 1. For each i, we simulate $X^{(i)}$ having the exponential distribution with parameter $\mu^{(i)}$. We treat $X^{(1)},\dots,X^{(10,000)}$ as $k=20$ samples of size $m=500$ each, and we compute the sample median $Z_1,\dots,Z_{20}$ of each of the 20 samples. After performing these initial simulations, suppose that we observe the value $S=0.01597$ for Eq. (12.2.2). This makes $\hat{\sigma}=0.3570$. Plugging this value into (12.2.5) for σ with $\gamma=0.99$ and $\epsilon=0.001$ yields $v=845{,}747.4$. This means that we need a total of 845,748 simulations to reach our desired level of confidence in the simulated result. Just to check, we simulated a total of 900,000 values and computed the sample median 0.2593 as well as a new value of S^2 based on $k=100$ subsamples of size $m=6200$ each. The new value of $\hat{\sigma}$ is 0.4529. Substituting 0.4529 for σ in Eq. (12.2.5) yields a new v =1,360,939, which means that we still need another 460,939 simulations. ◀

Simulating Real Processes

In many scientific fields, real physical or social processes are modeled as having random components. For example, stock prices are often modeled as having lognormal distributions as in Example 5.6.10. Many processes involving waiting times and service are modeled using Poisson processes. The simple probability models that have been developed earlier in this text are merely the building blocks of which such models of real processes are constructed. Here, we shall give two examples of slightly more complicated models that can be constructed using the distributions we already know. The analyses of these models can be simplified by the use of simulation.

Example 12.2.13 Option Pricing. In Example 5.6.10, we introduced the formula of Black and Scholes (1973) for pricing options. In that example, the option was to buy shares at price q of a stock whose value at time u (in the future) is a random variable S_u with a known lognormal distribution. Many financial analysts believe that the standard deviation σ of $\log(S_u)$ in Example 5.6.10 should not be treated as a known constant. For example, we could treat σ as a random variable with a p.d.f. $f(\sigma)$. To be precise, we shall continue to assume that $S_u=S_0e^{(r-\sigma^2/2)u+\sigma u^{1/2}Z}$, but now we shall assume that both Z and σ are random variables. For convenience, we shall assume that they are independent. We shall let Z have the standard normal distribution, and we shall let $\tau=1/\sigma^2$ have the gamma distribution with known parameters α and β. The parameters α and β might result from estimating the variance of stock prices based on historical data combined with expert opinion of stock analysts. For example,

they might be the posterior hyperparameters that result from applying a Bayesian analysis to a sample of stock prices. It is easy to see that $E(S_u|\sigma) = S_0 e^{ru}$ for all σ, and hence the law of total probability for expectations (Theorem 4.7.1) implies that $E(S_u) = S_0 e^{ru}$. This is what we need for risk neutrality. The price for the option considered in Example 5.6.10 is the mean of the random variable $e^{-ru}h(S_u)$, where

$$h(s) = \begin{cases} s - q & \text{if } s > q, \\ 0 & \text{otherwise.} \end{cases}$$

The Black-Scholes formula (5.6.18) is just the conditional mean of $e^{-ru}h(S_u)$ given σ. To estimate the marginal mean of $e^{-ru}h(S_u)$, we could simulate a large number of values $\sigma^{(i)}$ $(i = 1, \dots, v)$ from the distribution of σ, substitute each $\sigma^{(i)}$ into (5.6.18), and average the results.

As an example, suppose that we take the same numerical situation from the end of Example 5.6.10 with $u = 1$, $r = 0.06$, and $q = S_0$. This time, suppose that $1/\sigma^2$ has the gamma distribution with parameters 2 and 0.0127. (These numbers make $E(\sigma) = 0.1$, but σ has substantial variability.) We can sample v =1,000,000 values of σ from this distribution and compute (5.6.18) for each value. The average, in our simulation, is $0.0756 S_0$, and the simulation standard error is $1.814 S_0 \times 10^{-5}$. The option price is only slightly higher than it was when we assumed that we knew σ. When the distribution of S_u is even more complicated, one can simulate S_u directly and estimate the mean of $h(S_u)$. ◀

In the following example, each simulation requires a large number of steps, but each step is relatively simple. The combination of several simple steps into one complicated step is very common in simulations of real processes.

Example 12.2.14

A Service Queue with Impatient Customers. Consider a queue to which customers arrive according to a Poisson process with rate λ per hour. Suppose that the queue has a single server. Each customer who arrives at the queue counts the length r of the queue (including the customer being served) and decides to leave with probability p_r, for $r = 1, 2, \dots$. A customer who leaves does *not* enter the queue. Each customer who enters the queue waits in the order of arrival until the customer immediately in front is done being served, and then moves to the head of the queue. The time (in hours) to serve a customer, after reaching the head of the queue, is an exponential random variable with parameter μ. Assume that all service times are independent of each other and of all arrival times.

We can use simulation to learn about the behavior of such a queue. For example, we could estimate the expected number of customers in the queue at a particular time t after the queue opens for business. To do this, we could simulate many, say, v, realizations of the queue operation. For each realization i, we count how many customers $N^{(i)}$ are in the queue at time t. Then our estimator is $\frac{1}{v}\sum_{i=1}^{v} N^{(i)}$. To simulate a single realization, we could proceed as follows: Simulate interarrival times $X_1, X_2, \dots$ of the Poisson process as i.i.d. exponential random variables with parameter λ. Let $T_j = \sum_{i=1}^{j} X_i$ be the time at which customer j arrives. Stop simulating at the first k such that $T_k > t$. Only the first $k - 1$ customers have even arrived at the queue by time t. For each $j = 1, \dots, k - 1$, simulate a service time Y_j having the exponential distribution with parameter μ. Let Z_j stand for the time at which the jth customer reaches the head of the queue, and let W_j stand for the time at which the jth customer leaves the queue. For example, $Z_1 = X_1$ and $W_1 = X_1 + Y_1$. For $j > 1$, the jth customer first counts the length of the queue and decides whether or not to leave. Let $U_{i,j} = 1$ if customer i is still in the queue when customer j arrives $(i < j)$,

and let $U_{i,j} = 0$ if customer i has already left the queue. Then

$$U_{i,j} = \begin{cases} 1 & \text{if } W_i \geq T_j, \\ 0 & \text{otherwise.} \end{cases}$$

The number of customers in the queue when the jth customer arrives is $r = \sum_{i=1}^{j-1} U_{i,j}$. We then simulate a random variable V_j having the Bernoulli distribution with parameter p_r. If $V_j = 1$, customer j leaves the queue so that $W_j = T_j$. If customer j stays in the queue, then this customer reaches the head of the queue at time

$$Z_j = \max\{T_j, W_1, \ldots, W_{j-1}\}.$$

That is, the jth customer either reaches the head of the queue immediately upon arrival (if nobody is still being served) or as soon as all of the previous $j-1$ customers have left, whichever comes later. Also, $W_j = Z_j + Y_j$ if customer j stays. For each $j = 1, \ldots, k-1$, the jth customer is in the queue at time t if and only if $W_j \geq t$.

As a numerical example, suppose that $\lambda = 2$, $\mu = 1$, $t = 3$, and $p_r = 1 - 1/r$, for $r \geq 1$. Suppose that the first $k = 6$ simulated interarrival times are

$$0.215,\ 0.713,\ 1.44,\ 0.174,\ 0.342,\ 0.382.$$

The sum of the first five of these times is 2.884, but the sum of all six is 3.266. So, at most five customers are in the queue at time $t = 3$. Suppose that the simulated service times for the first five customers are

$$0.251,\ 2.215,\ 2.855,\ 0.666,\ 2.505.$$

We cannot simulate the V_j's in advance, because we do not yet know how many customers will be in the queue when each customer j arrives. Figure 12.3 shows a time line of the simulation of the process that we are about to describe. Begin with customer 1, who has $T_1 = Z_1 = 0.215$ and $W_1 = 0.215 + 0.251 = 0.466$. For customer 2, $T_2 = T_1 + 0.713 = 0.928 > W_1$, so nobody is in the queue when customer 2 arrives and $Z_2 = T_2 = 0.928$. Then $W_2 = Z_2 + 2.215 = 3.143$. For customer 3, $T_3 = T_2 + 1.44 = 2.368 < W_2$, so $r = 1$. Because $p_1 = 0$, customer 3 stays, and there is no need to

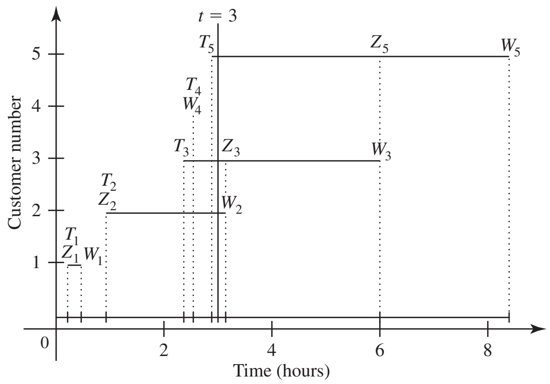


Figure 12.3 One simulation of a service queue. The bottom line is the time line for Example 12.2.14. Each customer is represented by one horizontal line segment. The vertical line at $t = 3$ crosses the horizontal lines for those customers still in the queue at time $t = 3$.

simulate V_3. Then $Z_3 = W_2 = 3.143$, and $W_3 = Z_3 + 2.855 = 5.998$. For customer 4, $T_4 = T_3 + 0.174 = 2.542$. Since $W_1 < T_4 < W_2, W_3$, we have $r = 2$ customers in the queue. We then simulate V_4 having the Bernoulli distribution with parameter $p_2 = 1/2$. Suppose that we simulate $V_4 = 1$, so customer 4 leaves, and we ignore the fourth simulated service time. This makes $W_4 = T_4 = 2.542$. For customer 5, $T_5 = T_4 + 0.342 = 2.884$, and customers 2 and 3 are still in the queue. We need to simulate V_5 having the Bernoulli distribution with parameter $p_2 = 1/2$. Suppose that $V_5 = 0$, so customer 5 stays. Then $Z_5 = W_3 = 5.988$, and $W_5 = Z_5 + 2.505 = 8.393$. Finally, $W_j \geq 3$ for $j = 2, 3, 5$. This means that there are $N^{(1)} = 3$ customers in the queue at time $t = 3$, as illustrated in Fig. 12.3. Needless to say, a computer should be programmed to do this calculation for a large simulation. ◄

Summary

If we wish to compute the expected value θ of some random variable Y, but cannot perform the necessary calculation in closed form, we can use simulation. In general, we would simulate a large random sample $Y^{(1)}, \dots, Y^{(v)}$ from the same distribution as Y, and then compute the sample mean Z as our estimator. We can also estimate a quantile θ_p of a distribution in a similar fashion. If $Y^{(1)}, \dots, Y^{(v)}$ is a large sample from the distribution, we can compute the sample p quantile Z. It is always a good idea to compute some measure of how good a simulation estimator is. One common measure is the simulation standard error of Z, an estimate of the standard deviation of the simulation distribution of Z. Alternatively, one could perform enough simulations to make sure that the probability is high that the Z is close to the parameter being estimated.

Exercises

1. Eq. (12.2.4) is based on the assumption that Z has approximately a normal distribution. Occasionally, the normal approximation is not good enough. In such cases, one can let

$$v = \frac{\sigma^2}{\epsilon^2(1-\gamma)}. \qquad (12.2.6)$$

To be precise, let Z be the average of v independent random variables with mean μ and variance σ^2. Prove that if v is at least as large as the number in Eq. (12.2.6), then $\Pr(|Z - \mu| \leq c) \geq \gamma$. *Hint:* Use the Chebyshev inequality (6.2.3).

2. In Example 12.2.11, how large would v need to be according to Eq. (12.2.6)?

3. Suppose that we have available as many i.i.d. standard normal random variables as we desire. Let X stand for a random variable having the normal distribution with mean 2 and variance 49. Describe a method for estimating $E(\log(|X| + 1))$ using simulation.

4. Use a pseudo-random number generator to simulate a sample of 15 independent observations in which 13 of the 15 are drawn from the uniform distribution on the interval $[-1, 1]$ and the other two are drawn from the uniform distribution on the interval $[-10, 10]$. For the 15 values that are obtained, calculate the values of (a) the sample mean, (b) the trimmed means for $k = 1, 2, 3$, and 4 (see Sec. 10.7), and (c) the sample median. Which of these estimators is closest to 0?

5. Repeat Exercise 4 ten times, using a different pseudo-random sample each time. In other words, construct 10 independent samples, each of which contains 15 observations and each of which satisfies the conditions of Exercise 4.

a. For each sample, which of the estimators listed in Exercise 4 is closest to 0?

b. For each of the estimators listed in Exercise 4, determine the square of the distance between the estimator and 0 in each of the 10 samples, and determine the average of these 10 squared distances. For which of the estimators is this average squared distance from 0 smallest?

6. Suppose that X and Y are independent, that X has the beta distribution with parameters 3.5 and 2.7, and that Y has the beta distribution with parameters 1.8 and 4.2. We are interested in the mean of $X/(X+Y)$. You may assume that you have the ability to simulate as many random variables with whatever beta distributions you wish.

a. Describe a simulation plan that will produce a good estimator of the mean of $X/(X+Y)$ if enough simulations are performed.

b. Suppose that you want to be 98 percent confident that your estimator is no more than 0.01 away from the actual value of $E[X/(X+Y)]$. Describe how you would determine an appropriate size for the simulation.

7. Consider the numbers in Table 10.40 on page 676. Suppose that you have available as many standard normal random variables and as many uniform random variables on the interval [0, 1] as you desire. You want to perform a simulation to obtain the number in the "Sample median" row and $\epsilon = 0.05$ column.

a. Describe how to perform such a simulation. *Hint:* Let X and U be independent such that X has the standard normal distribution and U has the uniform distribution on the interval [0, 1]. Let $0 < \epsilon < 1$, and find the distribution of

$$Y = \begin{cases} X & \text{if } U > \epsilon, \\ 10X & \text{if } U < \epsilon. \end{cases}$$

b. Perform the simulation on a computer.

8. Consider the same situation described in Exercise 7. This time, consider the number in the "Trimmed mean for $k = 2$" row and $\epsilon = 0.1$ column.

a. Describe how to perform a simulation to produce this number.

b. Perform the simulation on a computer.

9. In Example 12.2.12, we can actually compute the median θ of the distribution of the X_i in closed form. Calculate the true median, and see how far the simulated value was from the true value. *Hint:* Find the marginal p.d.f. of X by using the law of total probability for random variables (3.6.12) together with Eq. (5.7.10). The c.d.f. and quantile function are then easy to derive.

10. Let $X_1, \dots, X_{21}$ be i.i.d. with the exponential distribution that has parameter λ. Let M stand for the sample median. We wish to compute the M.S.E. of M as an estimator of the median of the distribution of the X_i's.

a. Determine the median of the distribution of X_1.

b. Let θ be the M.S.E. of the sample median when $\lambda = 1$. Prove that the M.S.E. of the sample median equals θ/λ^2 in general.

c. Describe a simulation method for estimating θ.

11. In Example 12.2.4, there is a slightly simpler way to simulate a sample from the posterior distribution of $\mu_x - \mu_y$. Suppose that we can simulate as many independent t pseudo-random variables as we wish with whatever degrees of freedom we want. Explain how we could use these t random variables to simulate a sample from the posterior distribution of $\mu_x - \mu_y$.

12. Let $(Y_1, W_1), \dots, (Y_n, W_n)$ be an i.i.d. sample of random vectors with finite covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{yy} & \sigma_{yw} \\ \sigma_{yw} & \sigma_{ww} \end{pmatrix}.$$

Let $\overline{Y}$ and $\overline{W}$ be the sample averages. Let $g(y, w)$ be a function with continuous partial derivatives g_1 and g_2 with respect to y and w, respectively. Let $Z = g(\overline{Y}, \overline{W})$. The two-dimensional Taylor expansion of g around a point (y_0, w_0) is

$$g(y, w) = g(y_0, w_0) + g_1(y_0, w_0)(y - y_0) + g_2(y_0, w_0)(w - w_0), \qquad (12.2.7)$$

plus an error term that we shall ignore here. Let $(y, w) = (\overline{Y}, \overline{W})$ and $(y_0, w_0) = (E(Y), E(W))$ in Eq. (12.2.7). To the level of approximation of Eq. (12.2.7), prove that

$$\begin{aligned}\text{Var}(Z) = {} & g_1(E(Y), E(W))^2\sigma_{yy} \\ & + 2g_1(E(Y), E(W))g_2(E(Y), E(W))\sigma_{yw} \\ & + g_2(E(Y), E(W))^2\sigma_{ww}.\end{aligned}$$

Hint: Use the formula for the variance of a linear combination of random variables derived in Sec. 4.6.

13. Use the two-dimensional delta method from Exercise 12 to derive the estimator of the simulation variance of a sample variance as given in Eq. (12.2.3). *Hint:* Replace $E(Y)$ and $E(W)$ by $\overline{Y}$ and $\overline{W}$, respectively, and replace $\mathbf{\Sigma}$ by the sample variances and sample covariance.

14. Let Y be a random variable with some distribution. Suppose that you have available as many pseudo-random variables as you want with the same distribution as Y. Describe a simulation method for estimating the skewness of the distribution of Y. (See Definition 4.4.1.)

15. Suppose that the price of a stock at time u in the future is a random variable $S_u = S_0 e^{\alpha u + W_u}$, where S_0 is the current price, α is a constant, and W_u is a random variable with known distribution. Suppose that you have available as many i.i.d. random variables as you wish with the distribution of W_u. Suppose that the m.g.f. $\psi(t)$ of W_u is known and finite on an interval that contains $t = 1$.

a. What number should α equal in order that $E(S_u) = e^{ru}S_0$?

b. We wish to price an option to purchase one share of this stock at time u for the price q. Describe how you could use simulation to estimate the price of such an option.

16. Consider a queue to which customers arrive according to a Poisson process with rate λ per hour. Suppose that the queue has two servers. Each customer who arrives at the queue counts the length r of the queue (including any customers being served) and decides to leave with probability p_r, for $r = 2, 3, \dots$. A customer who leaves does *not* enter the queue. Each customer who enters the queue waits in the order of arrival until at least one of the two servers is available, and then begins being served by the available server. If both servers are available, the customer chooses randomly between the two servers with probability 1/2 for each, independent of all other random variables. For server i ($i = 1, 2$), the time (in hours) to serve a customer, after beginning service, is an exponential random variable with parameter μ_i. Assume that all service times are independent of each other and of all arrival times. Describe how to simulate the number of customers in the queue (including any being served) at a specific time t.

12.3 Simulating Specific Distributions

In order to perform statistical simulations, we must be able to obtain pseudo-random values from a variety of distributions. In this section, we introduce some methods for simulating from specific distributions.

Most computer packages with statistical capability are able to generate pseudo-random numbers with the uniform distribution on the interval [0, 1]. We shall assume throughout the remainder of this section that one has available an arbitrarily large sample of what appear to be i.i.d. random variables (pseudo-random numbers) with the uniform distribution on the interval [0, 1]. Usually, we need random variables with other distributions, and the purpose of this section is to review some common methods for transforming uniform random variables into random variables with other distributions.

The Probability Integral Transformation

In Chapter 3, we introduced the probability integral transformation for transforming a uniform random variable X on the interval [0, 1] into a random variable Y with a continuous strictly increasing c.d.f. G. The method is to set $Y = G^{-1}(X)$. This method works well if G^{-1} is easily computed.

Example 12.3.1 Generating Exponential Pseudo-Random Variables. Suppose that we want Y to have the exponential distribution with parameter λ, where λ is a known constant. The c.d.f. of Y is

$$G(y) = \begin{cases} 1 - e^{-\lambda y} & \text{if } y \ge 0, \\ 0 & \text{if } y < 0. \end{cases}$$

We can easily invert this function to obtain

$$G^{-1}(x) = -\log(1 - x)/\lambda, \quad \text{if } 0 < x < 1.$$

If X has the uniform distribution on the interval [0, 1], then $-\log(1 - X)/\lambda$ has the exponential distribution with parameter λ. ◀

Special-Purpose Algorithms

There are cases in which the desired c.d.f. G is not easy to invert. For example, if G is the standard normal c.d.f., then G^{-1} must be obtained by numerical approximation.

However, there is a clever method for transforming two independent uniform random variables on the interval [0, 1] into two standard normal random variables. The method was described by Box and Müller (1958).

Example 12.3.2 Generating Two Independent Standard Normal Variables. Let X_1, X_2 be independent with the uniform distribution on the interval [0, 1]. The joint p.d.f. of (X_1, X_2) is

$$f(x_1, x_2) = 1, \quad \text{for } 0 < x_1, x_2 < 1.$$

Define

$$Y_1 = [-2\log(X_1)]^{1/2}\sin(2\pi X_2),$$
$$Y_2 = [-2\log(X_1)]^{1/2}\cos(2\pi X_2).$$

The inverse of this transformation is

$$X_1 = \exp[-(Y_1^2 + Y_2^2)/2],$$
$$X_2 = \frac{1}{2\pi}\arctan(Y_1/Y_2).$$

Using the methods of Sec. 3.9, we compute the Jacobian, which is the determinant of the matrix of partial derivatives of the inverse function:

$$\begin{pmatrix} -y_1\exp[-(y_1^2 + y_2^2)/2] & -y_2\exp[-(y_1^2 + y_2^2)/2] \\ \frac{1}{2\pi y_2}\frac{1}{1+(y_1/y_2)^2} & -\frac{y_1}{2\pi y_2^2}\frac{1}{1+(y_1/y_2)^2} \end{pmatrix}.$$

The determinant of this matrix is $J = \exp[-(y_1^2 + y_2^2)/2]/(2\pi)$. The joint p.d.f. of (Y_1, Y_2) is then

$$g(y_1, y_2) = f\Big(\exp[(y_1^2 + y_2^2)/2], \arctan(y_1/y_2)/(2\pi)\Big)\,|J|$$
$$= \exp[-(y_1^2 + y_2^2)/2]/(2\pi).$$

This is the joint p.d.f. of two independent standard normal variables. ◂

Acceptance/Rejection

Many other special-purpose methods exist for other distributions, also. We would like to present here one more general-purpose method that has wide applicability. The method is called *acceptance/rejection*. Let f be a p.d.f. and assume that we would like to sample a pseudo-random variable with this p.d.f. Assume that there exists another p.d.f. g with the following two properties:

- We know how to simulate a pseudo-random variable with p.d.f. g.
- There exists a constant k such that $kg(x) \ge f(x)$ for all x.

To simulate a single Y with p.d.f. f, perform the following steps:

1. Simulate a pseudo-random X with p.d.f. g and an independent uniform pseudo-random variable U on the interval [0, 1].
2. If
$$\frac{f(X)}{g(X)} \ge kU, \tag{12.3.1}$$
let $Y = X$, and stop the process.
3. If (12.3.1) fails, throw away X and U, and return to the first step.

If we need more than one Y, we repeat the entire process as often as needed. We now show that the p.d.f. of each Y is f.

Theorem 12.3.1 The p.d.f. of Y in the acceptance/rejection method is f.

Proof First, we note that the distribution of Y is the conditional distribution of X given that (12.3.1) holds. That is, let A be the event that (12.3.1) holds, and let $h(x, u|A)$ be the conditional joint p.d.f. of (X, U) given A. Then the p.d.f. of Y is $\int h(x, u|A)\,du$. This is because Y is constructed to be X conditional on (12.3.1) holding. The conditional p.d.f. of (X, U) given A is

$$h(x, u|A) = \frac{1}{\Pr(A)} \begin{cases} g(x) & \text{if } f(x)/g(x) \ge ku \text{ and } 0 < u < 1, \\ 0 & \text{otherwise.} \end{cases}$$

It is straightforward to calculate $\Pr(A)$, that is, the probability that $U \le f(X)/[kg(X)]$.

$$\Pr(A) = \int_{-\infty}^{\infty} \int_{0}^{f(x)/[kg(x)]} g(x)\,du\,dx = \int_{-\infty}^{\infty} \frac{1}{k} f(x)\,dx = \frac{1}{k}.$$

So,

$$h(x, u|A) = k \begin{cases} g(x) & \text{if } f(x)/g(x) \ge ku \text{ and } 0 < u < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The integral of this function over all u values for fixed x is the p.d.f. of Y evaluated at x:

$$\int h(x, u|A)\,du = k \int_{0}^{f(x)/[kg(x)]} g(x)\,du = f(x). \qquad \blacksquare$$

Here is an example of the use of acceptance/rejection.

Example 12.3.3 Simulating a Beta Distribution. Suppose that we wish to simulate a random variable Y having the beta distribution with parameters 1/2 and 1/2. The p.d.f. of Y is

$$f(y) = \frac{1}{\pi} y^{-1/2}(1-y)^{-1/2}, \text{ for } 0 < y < 1.$$

Note that this p.d.f. is unbounded. However, it is easy to see that

$$f(y) \le \frac{1}{\pi}(y^{-1/2} + (1-y)^{-1/2}), \tag{12.3.2}$$

for all $0 < y < 1$. The right side of Eq. (12.3.2) can be written as $kg(y)$ with $k = 4/\pi$ and

$$g(y) = \frac{1}{2}\left[\frac{1}{2y^{1/2}} + \frac{1}{2(1-y)^{1/2}}\right].$$

This g is a half-and-half mixture of two p.d.f.'s g_1 and g_2:

$$g_1(x) = \frac{1}{2x^{1/2}}, \text{ for } 0 < x < 1,$$

$$g_2(x) = \frac{1}{2(1-x)^{1/2}}, \text{ for } 0 < x < 1. \tag{12.3.3}$$

We can easily simulate random variables from these distributions using the probability integral transformation. To simulate a random variable X with p.d.f. g, simulate three random independent variables U_1, U_2, U_3 with uniform distributions on the

interval [0, 1]. If $U_1 \le 1/2$, simulate X from g_1 using the probability integral transformation applied to U_2. If $U_1 > 1/2$, simulate X from g_2 using the probability integral transformation and U_2. If $f(X)/g(X) \ge kU_3$, let $Y = X$. If not, repeat the process. ◄

When using the acceptance/rejection method, one must usually reject simulated values and resimulate. The probability of accepting a value is $\Pr(A)$ in the proof of Theorem 12.3.1, namely, $1/k$. The larger k is, the harder it will be to accept. In Exercise 5, you will prove that the expected number of iterations until the first acceptance is k.

A common special case of acceptance/rejection is the simulation of a random variable conditional on some event. For example, let X be a random variable with the p.d.f. g, and suppose that we want the conditional distribution of X given that $X > 2$. Then the conditional p.d.f. of X given $X > 2$ is

$$f(x) = \begin{cases} kg(x) & \text{if } x > 2, \\ 0 & \text{if } x \le 2, \end{cases}$$

where $k = 1/\int_2^\infty g(x)\,dx$. Note that $f(x) \le kg(x)$ for all x, so acceptance/rejection is applicable. In fact, since $f(X)/g(X)$ only takes the two values k and 0, we don't need to simulate the uniform U in the acceptance/rejection algorithm. We don't even need to compute the value k. We just reject each $X \le 2$. Here is a version of the same algorithm to solve a question that was left open in Sec. 11.8.

Example 12.3.4

Computing the Size of a Two-Stage Test. In Sec. 11.8, we studied the analysis of data from a two-way layout with replication. In that section, we introduced a two-stage testing procedure. First, we tested the hypotheses (11.8.11), and then, if we accepted the null hypothesis, we proceeded to test the hypotheses (11.8.13). Unfortunately, we were unable to compute the conditional size of the second test given that the first test accepted the null hypothesis. That is, we could not calculate (11.8.15) in closed form. However, we can use simulation to estimate the conditional size.

The two tests are based on U_{AB}^2, defined in Eq. (11.8.12), and V_A^2, defined in Eq. (11.8.16). The first test rejects the null hypothesis in (11.8.11) if $U_{AB}^2 \ge d$, where d is a quantile of the appropriate F distribution. The second test rejects its null hypothesis if $V_A^2 \ge c$, where c is yet to be determined. The random variables U_{AB}^2 and V_A^2 are both ratios of various mean squares. In particular, they share a common denominator $MS_{\text{Resid}} = S^2_{\text{Resid}}/[IJ(K-1)]$. In order to determine an appropriate critical value c for the second test, we need the conditional distribution of V_A^2 given that $U_{AB}^2 < d$, and given that both null hypotheses are true. We can sample from that conditional distribution as follows: Let the interaction mean square be $MS_{AB} = S^2_{\text{Int}}/[(I-1)(J-1)]$, and let the mean square for factor A be $MS_A = S_A^2/(I-1)$. Then $U_{AB}^2 = MS_{AB}/MS_{\text{Resid}}$ and $V_A^2 = MS_A/MS_{\text{Resid}}$. All of these mean squares are independent, and they all have different gamma distributions when the null hypotheses are both true. Most statistical computer packages will allow simulation of gamma random variables. So, we start by simulating many triples $(MS_{AB}, MS_{\text{Resid}}, MS_A)$. Then, for each simulated triple, we compute U_{AB}^2 and V_A^2. If $U_{AB}^2 \ge d$, we discard the corresponding V_A^2. The undiscarded V_A^2 values are a random sample from the conditional distribution that we need. The efficiency of this algorithm could be improved slightly by simulating MS_A and then computing V_A^2 only when $U_{AB}^2 < d$ is observed. ◄

Generating Functions of Other Random Variables

It often happens that there is more than one way to simulate from a particular distribution. For example, suppose that a distribution is defined as the distribution of a particular function of other random variables (in the way that the χ^2, t, and F distributions are). In such cases, there is a straightforward way to simulate the desired distribution. First, simulate the random variables in terms of which the distribution is defined, and then calculate the appropriate function.

Example 12.3.5 Alternate Method for Simulating a Beta Distribution. In Exercise 6 in Sec. 5.8, you proved the following: If U and V are independent, with U having the gamma distribution with parameters α_1 and β, and V having the gamma distribution with parameters α_2 and β, then $U/(U+V)$ has the beta distribution with parameters α_1 and α_2. So, if we have a method for simulating gamma random variables, we can simulate beta random variables. The case handled in Example 12.3.3 is $\alpha_1 = \alpha_2 = 1/2$. Let $\beta = 1/2$ so that U and V would both have gamma distributions with parameters 1/2 and 1/2, also known as the χ^2 distribution with one degree of freedom. If we simulate two independent standard normal random variables X_1, X_2 (for example, by the method of Example 12.3.2), then X_1^2 and X_2^2 are independent and have the χ^2 distribution with one degree of freedom. It follows that $Y = X_1^2/(X_1^2 + X_2^2)$ has the beta distribution with parameters 1/2 and 1/2. ◀

As another example, to simulate a χ^2 random variable with 10 degrees of freedom, one could simulate 10 i.i.d. standard normals, square them, and add up the squares. Alternatively, one could simulate five random variables having the exponential distribution with parameter 1/2 and add them up.

Example 12.3.6 Generating Pseudo-Random Bivariate Normal Vectors. Suppose that we wish to simulate a bivariate normal vector with the p.d.f. given in Eq. (5.10.2). This p.d.f. was constructed as the joint p.d.f. of

$$\begin{aligned} X_1 &= \sigma_1 Z_1 + \mu_1, \\ X_2 &= \sigma_2\left[\rho Z_1 + (1-\rho^2)^{1/2} Z_2\right] + \mu_2, \end{aligned} \tag{12.3.4}$$

where Z_1 and Z_2 are i.i.d. with the standard normal distribution. If we use the method of Example 12.3.2 to generate independent Z_1 and Z_2 with the standard normal distribution, we can use the formulas in (12.3.4) to transform these into X_1 and X_2, which will then have the desired bivariate normal distribution. ◀

Most statistical computer packages have the capability of simulating pseudo-random variables with each of the continuous distributions that have been named in this text. The techniques of this section are really needed only for simulating less common distributions or when a statistical package is not available.

Some Examples Involving Simulation of Common Distributions

Example 12.3.7 Bayesian Analysis of One-Way Layout. We can perform a Bayesian analysis of a one-way layout using the same statistical model presented in Sec. 11.6 together with an improper prior for the model parameters. (We could use a proper prior, but the additional calculations would divert our attention from the simulation issues.) Let $\tau = 1/\sigma^2$, as we did in Sec. 8.6. The usual improper prior for the parameters $(\mu_1, \ldots, \mu_p, \tau)$ has "p.d.f." $1/\tau$. The posterior joint p.d.f. is then proportional to $1/\tau$

times the likelihood. The observed data are y_{ij} for $j = 1, \ldots, n_i$ and $i = 1, \ldots, p$. The likelihood function is

$$(2\pi)^{-n/2}\tau^{n/2}\exp\left(-\frac{\tau}{2}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2\right),$$

where $n = n_1 + \cdots + n_p$. To simplify the likelihood function, we can rewrite the sum of squares that appears in the exponent as

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2 = \sum_{i=1}^{p} n_i(\overline{y}_{i+}-\mu_i)^2 + S^2_{\text{Resid}},$$

where $\overline{y}_{i+}$ is the average of $y_{i1}, \ldots, y_{in_i}$ and

$$S^2_{\text{Resid}} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i+})^2$$

is the residual sum of squares. Then, the posterior p.d.f. is proportional to

$$\tau^{p/2}\exp\left(-\frac{\tau}{2}\sum_{i=1}^{p} n_i(\overline{y}_{i+}-\mu_i)^2\right)\tau^{(n-p)/2-1}\exp\left(-\frac{\tau}{2}S^2_{\text{Resid}}\right).$$

This expression is easily recognized as the product of the gamma p.d.f. for τ with parameters $(n-p)/2$ and $S^2_{\text{Resid}}/2$ and the product of p normal p.d.f.'s for $\mu_1, \ldots, \mu_p$ with means $\overline{y}_{i+}$ and precisions $n_i\tau$ for $i = 1, \ldots, p$. Hence, the posterior joint distribution of the parameters is the following: Conditional on τ, the μ_i's are independent with μ_i having the normal distribution with mean $\overline{y}_{i+}$ and precision $n_i\tau$. The marginal distribution of τ is the gamma distribution with parameters $(n-p)/2$ and $S^2_{\text{Resid}}/2$.

If we simulate a large sample of parameters from the posterior distribution, we could begin to answer questions about what we have learned from the data. To do this, we would first simulate a large number of τ values $\tau^{(1)}, \ldots, \tau^{(v)}$. Most statistical programs allow the user to simulate gamma random variables with arbitrary first parameter and second parameter 1. So, we could simulate $T^{(1)}, \ldots, T^{(v)}$ having the gamma distribution with parameters $(n-p)/2$ and 1. We could then let $\tau^{(\ell)} = 2T^{(\ell)}/S^2_{\text{Resid}}$ for $\ell = 1, \ldots, v$. Then, for each ℓ simulate independent $\mu_1^{(\ell)}, \ldots, \mu_p^{(\ell)}$ with $\mu_i^{(\ell)}$ having the normal distribution with mean $\overline{y}_{i+}$ and variance $1/[n_i\tau^{(\ell)}]$.

As a specific example, consider the hot dog data in Example 11.6.2. We begin by simulating $v = 60{,}000$ sets of parameters as described above. Now we can address the question of how much difference there is between the means. There are several ways to do this. We could compute the probability that all $|\mu_i - \mu_j| > c$ for each positive c. We could compute the probability that at least one $|\mu_i - \mu_j| > c$ for each positive c. We could compute the quantiles of $\max_{i,j}|\mu_i - \mu_j|$, of $\min_{i,j}|\mu_i - \mu_j|$, or of the average of all $|\mu_i - \mu_j|$. For example, in 99 percent of the 60,000 simulations, at least one $|\mu_i^{(\ell)} - \mu_j^{(\ell)}| > 27.94$. The simulation standard error of this estimator of the 0.99 quantile of $\max_{i,j}|\mu_i - \mu_j|$ is 0.1117. (For the remainder of this example, we shall present only the simulation estimates and not their simulation standard errors.) In about 1/2 of the simulations, all $|\mu_i^{(\ell)} - \mu_j^{(\ell)}| > 2.379$. And in 99 percent of the simulations, the average of the differences was at least 14.59. Whether 27.94, 14.59, or 2.379 count as large differences depends on what decisions we need to make about the hot dogs. A useful way to summarize all of these calculations is through a plot of the sample c.d.f.'s of the largest, smallest, and average of the six $|\mu_i - \mu_j|$ differences. (The sample c.d.f. of a set of numbers is defined at the very beginning of Sec. 10.6.)

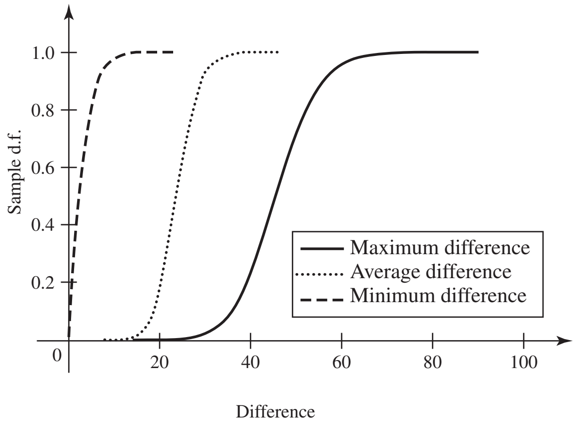


Figure 12.4 Sample c.d.f.'s of the maximum, average, and minimum of the six $|\mu_i - \mu_j|$ differences for Example 12.3.7.

Table 12.4 Posterior probabilities that each μ_i is largest and smallest in Example 12.3.7

Type	Beef	Meat	Poultry	Specialty
i	1	2	3	4
$\Pr(\mu_i \text{ largest}\mid \boldsymbol{y})$	0.1966	0.3211	0	0.4823
$\Pr(\mu_i \text{ smallest}\mid \boldsymbol{y})$	0	0	1	0

Figure 12.4 contains such a plot for this example. If we are simply concerned with whether or not there are any differences at all between the four types of hot dogs, then the "Maximum" curve in Fig. 12.4 is the one to examine. (Can you explain why this is the case?)

We can also attempt to answer questions that we would have great difficulty addressing in the ANOVA framework of Chapter 11. For example, we could ask what is the probability that each μ_i is the largest or smallest of the four. For each i, let N_i be the number of simulations j such that $\mu_i^{(j)}$ is the smallest of $\mu_1^{(j)}, \dots, \mu_4^{(j)}$. Also let M_i be the number of simulations j such that $\mu_i^{(j)}$ is the largest of the four means. Then $N_i/60{,}000$ is our simulation estimate of the probability that μ_i is the smallest mean, and $M_i/60{,}000$ is our estimate of the probability that μ_i is the largest mean. The results are summarized in Table 12.4. We see that μ_3 is almost certainly the smallest, while μ_4 has almost a 50 percent chance of being the largest. ◄

Example 12.3.8 Comparing Copper Ores. We shall illustrate the method of Example 12.2.4 using the data on copper ores from Example 9.6.5. Suppose that the prior distributions for all parameters are improper. The observed data consist of one sample of size 8 and another sample of size 10 with $\overline{X} = 2.6$, $\sum_{i=1}^{8}(X_i - \overline{X})^2 = 0.32$, $\overline{Y} = 2.3$, and $\sum_{j=1}^{10}(Y_j - \overline{Y})^2 = 0.22$. The posterior distributions then have hyperparameters $\mu_{x1} = 2.6$, $\lambda_{x1} = 8$, $\alpha_{x1} = 3.5$, $\beta_{x1} = 0.16$, $\mu_{y1} = 1.15$, $\lambda_{y1} = 10$, $\alpha_{y1} = 4.5$, and $\beta_{y1} = 0.11$. The posterior distributions of τ_x and τ_y are, respectively, the gamma distribution with parameters 3.5 and 0.16 and the gamma distribution with parameters 4.5 and

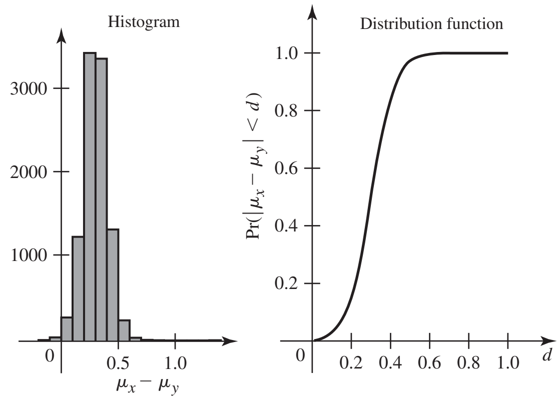


Figure 12.5 Histogram of simulated $\mu_x - \mu_y$ values together with posterior c.d.f. of $|\mu_x - \mu_y|$ for Example 12.3.8.

0.11. We can easily simulate, say, 10,000 pseudo-random values from each of these two distributions. For each simulated τ_x, we simulate a μ_x that has the normal distribution with mean 2.6 and variance $1/(8\tau_x)$. For each simulated τ_y, we simulate a μ_y that has the normal distribution with mean 2.3 and variance $1/(10\tau_y)$. Figure 12.5 contains a histogram of the 10,000 simulated $\mu_x - \mu_y$ values together with the sample c.d.f. of $|\mu_x - \mu_y|$. It appears that $\mu_x - \mu_y$ is almost always positive; indeed, it was positive for over 99 percent of the sampled values. The probability is quite high that $|\mu_x - \mu_y| < 0.5$, so that if 0.5 is not a large difference in this problem, we can be confident that μ_x and μ_y are pretty close. On the other hand, if 0.1 is a large difference, we can be confident that μ_x and μ_y are pretty far apart. ◀

If all we care about in Example 12.3.8 is the distribution of $\mu_x - \mu_y$, then we could simulate μ_x and μ_y directly without first simulating τ_x and τ_y. Since μ_x and μ_y are independent in this example, we could simulate each of them from their respective marginal distributions.

Example 12.3.9

Power of the *t* Test. In Theorem 9.5.3, we showed how the power function of the t test can be computed from the noncentral t distribution function. Not all statistical packages compute noncentral t probabilities. We can use simulation to estimate these probabilities. Let Y have the noncentral t distribution with m degrees of freedom and noncentrality parameter ψ. Then Y has the distribution of $X_1/(X_2/m)^{1/2}$ where X_1 and X_2 are independent with X_1 having the normal distribution with mean ψ and variance 1 and X_2 having the χ^2 distribution with m degrees of freedom. A simple way to estimate the c.d.f. of Y is to simulate a large number of (X_1, X_2) pairs and compute the sample c.d.f. of the values of $X_1/(X_2/m)^{1/2}$. ◀

The Simulation Standard Error of a Sample c.d.f In Examples 12.3.7 and 12.3.8, we plotted the sample c.d.f.'s of functions of simulated data. We did not associate simulation standard errors with these functions. We could compute simulation standard errors for every value of the sample c.d.f., but there is a simpler way to summarize the uncertainty about a sample c.d.f. We can make use of the Glivenko-Cantelli lemma (Theorem 10.6.1). To summarize that result in the context of simulation, let $Y^{(i)}$, $(i = 1, \ldots, v)$ be a simulated i.i.d. sample with c.d.f. G. Let G_v be the sample c.d.f. For each real x, $G_v(x)$ is the proportion of the simulated sample that is less than or equal to x. That is, $G_v(x)$ is $1/v$ times the number of i's such that $Y^{(i)} \le x$.

Theorem 10.6.1 says that if v is large, then

$$\Pr\left(|G_v(x) - G(x)| \le \frac{t}{v^{1/2}},\ \text{for all } x\right) \approx H(t),$$

where H is the function in Table 10.32 on page 661. In particular, with $t = 2$, $H(t) = 0.9993$. So we can declare (at least approximately) that $|G_v(x) - G(x)| \le 2/v^{1/2}$ simultaneously for all x with probability 0.9993. In Example 12.3.7, we had $v = 60,000$, so each curve in Fig. 12.4 should be accurate to within 0.008 with probability 0.9993. Indeed, all three curves simultaneously should be accurate to within 0.008 with probability 0.9979. (Prove this in Exercise 14.)

Simulating a Discrete Random Variable

All of the examples so far in this section have concerned simulations of random variables with continuous distributions. Occasionally, one needs random variables with discrete distributions. Algorithms for simulating discrete random variables exist, and we shall describe some here.

Example 12.3.10 **Simulating a Bernoulli Random Variable.** It is simple to simulate a pseudo-random Bernoulli random variable X with parameter p. Start with U having the uniform distribution on the interval [0, 1], and let $X = 1$ if $U \le p$. Otherwise, let $X = 0$. Since $\Pr(U \le p) = p$, X has the correct distribution. This method can be used to simulate from any distribution that is supported on only two values. If

$$f(x) = \begin{cases} p & \text{if } x = t_1, \\ 1 - p & \text{if } x = t_2, \\ 0 & \text{otherwise,} \end{cases}$$

then let $X = t_1$ is $U \le p$, and let $X = t_2$ otherwise. ◄

Example 12.3.11 **Simulating a Discrete Uniform Random Variable.** Suppose that we wish to simulate pseudo-random variables from a distribution that has the p.f.

$$f(x) = \begin{cases} \frac{1}{n} & \text{if } x \in \{t_1, \ldots, t_n\}, \\ 0 & \text{otherwise.} \end{cases} \tag{12.3.5}$$

The uniform distribution on the integers $1, \ldots, n$ is an example of such a distribution. A simple way to simulate a random variable with the p.f. (12.3.5) is the following: Let U have the uniform distribution on the interval [0, 1], and let Z be the greatest integer less than or equal to $nU + 1$. It is easy to see that Z takes the values $1, \ldots, n$ with equal probability, and so $X = t_Z$ has the p.f. (12.3.5). ◄

The method described in Example 12.3.11 does not apply to more general discrete distributions. However, the method of Example 12.3.11 is useful in simulations that are done in bootstrap analyses described in Sec. 12.6.

For general discrete distributions, there is an analog to the probability integral transformation. Suppose that a discrete distribution is concentrated on the values $t_1 < \cdots < t_n$ and that the c.d.f. is

$$F(x) = \begin{cases} 0 & \text{if } x < t_1, \\ q_i & \text{if } t_i \le x < t_{i+1}, \text{ for } i = 1, \ldots, n-1, \\ 1 & \text{if } x \ge t_n. \end{cases} \tag{12.3.6}$$

The following is the quantile function from Definition 3.3.2:

$$F^{-1}(p) = \begin{cases} t_1 & \text{if } 0 < p \le q_1, \\ t_{i+1} & \text{if } q_i < p \le q_{i+1}, \text{for } i = 1, \ldots, n-2, \\ t_n & \text{if } q_{n-1} < p < 1. \end{cases} \tag{12.3.7}$$

You can prove (see Exercise 13) that if U has the uniform distribution on the interval [0, 1], then $F^{-1}(U)$ has the c.d.f. in Eq. (12.3.6). This gives a straightforward, but inefficient, method for simulating arbitrary discrete distributions. Notice that the restriction that n be finite is not actually necessary. Even if the distribution has infinitely many possible values, F^{-1} can be defined by (12.3.7) by replacing $n-2$ by ∞ and removing the last branch.

Example 12.3.12 **Simulating a Geometric Random Variable.** Suppose that we wish to simulate a pseudo-random X having the geometric distribution with parameter p. In the notation of Eq. (12.3.7), $t_i = i-1$ for $i = 1, 2, \ldots$, and $q_i = 1-(1-p)^i$. Using the probability integral transformation, we would first simulate U with the uniform distribution on the interval [0, 1]. Then we would compare U to q_i for $i = 1, 2, \ldots$, until the first time that $q_i < U$ and set $X = i$. In this example, we can avoid the sequence of comparisons because we have a simple formula for q_i. The first i such that $q_i < U$ is the greatest integer strictly less than $\log(1-U)/\log(1-p)$. ◀

The probability integral transformation is very inefficient for discrete distributions that do not have a simple formula for q_i if the number of possible values is large. Walker (1974) and Kronmal and Peterson (1979) describe a more efficient method called the *alias method*. The alias method works as follows: Let f be the p.f. from which we wish to simulate a random variable X. Suppose that $f(x) > 0$ for only n different values of x. First, we write f as an average of n p.f.'s that are concentrated on one or two values each. That is,

$$f(x) = \frac{1}{n}[g_1(x) + \cdots + g_n(x)], \tag{12.3.8}$$

where each g_i is the p.f. of a distribution concentrated on one or two values only. We shall show how to do this in Example 12.3.13. To simulate X, first simulate an integer I that has the uniform distribution over the integers $1, \ldots, n$. (Use the method of Example 12.3.11.) Then simulate X from the distribution with the p.f. g_I. The reader can prove in Exercise 17 that X has the p.f. f.

Example 12.3.13 **Simulating a Binomial Random Variable Using the Alias Method.** Suppose that we need to simulate many random variables with a binomial distribution having parameters 9 and 0.4. The p.f. f of this distribution is given in a table at the end of this book. The distribution has $n = 10$ different values with positive probability. Since the n probabilities must add to 1, there must be x_1 and y_1 such that $f(x_1) \le 1/n$ and $f(y_1) \ge 1/n$. For example, $x_1 = 0$ and $y_1 = 2$ have $f(x_1) = 0.0101$ and $f(y_1) = 0.1612$. Define the first two-point p.f., g_1, as

$$g_1(x) = \begin{cases} nf(x_1) & \text{if } x = x_1, \\ 1 - nf(x_1) & \text{if } x = y_1, \\ 0 & \text{otherwise.} \end{cases}$$

In our case, $g_1(0) = 0.101$ and $g_1(2) = 0.899$. We then write f as $f(x) = g_1(x)/n + f_1^*(x)$, where

$$f_1^*(x) = \begin{cases} 0 & \text{if } x = x_1, \\ f(y_1) - g_1(y_1)/n & \text{if } x = y_1, \\ f(x) & \text{otherwise.} \end{cases}$$

In our example, $f_1^*(2) = 0.0713$. Now, f_1^* is positive at only $n-1$ different values, and the sum of the positive values of f_1^* is $(n-1)/n$. Hence, there must exist x_2 and y_2 such that $f_1^*(x_2) \le 1/n$ and $f_1^*(y_2) \ge 1/n$. For example, $x_2 = 2$ and $y_2 = 3$ have $f_1^*(x_2) = 0.0713$ and $f_1^*(y_2) = 0.2508$. Define g_2 by

$$g_2(x) = \begin{cases} nf_1^*(x_2) & \text{if } x = x_2, \\ 1 - nf_1^*(x_2) & \text{if } x = y_2, \\ 0 & \text{otherwise.} \end{cases}$$

Here, $g_2(2) = 0.713$. Now write $f_1^*(x) = g_2(x)/n + f_2^*(x)$, where

$$f_2^*(x) = \begin{cases} 0 & \text{if } x = x_2, \\ f_1^*(y_2) - g_2(y_2)/n & \text{if } x = y_2, \\ f_1^*(x) & \text{otherwise.} \end{cases}$$

In our example, $f_2^*(3) = 0.2221$. Now, f_2^* takes only $n-2$ positive values that add up to $(n-2)/n$. We can repeat this process $n-3$ more times, obtaining $g_1, \dots, g_{n-1}$ and f_{n-1}^*. Here, $f_{n-1}^*(x)$ takes only one positive value, at $x = x_n$, say, and $f_{n-1}^*(x_n) = 1/n$. Let g_n be a degenerate distribution at x_n. Then $f(x) = [g_1(x) + \cdots + g_n(x)]/n$ for all x.

After all of this initial setup, the alias method allows rapid simulation from f as follows: Simulate independent U and I with U having the uniform distribution on the interval [0, 1] and I having the uniform distribution on the integers $1, \dots, n$ ($n = 10$ in our example). If $U \le g_I(x_I)$, set $X = x_I$. If $U > g_I(x_I)$, set $X = y_I$. Here, the values we need to perform the simulation are

i	1	2	3	4	5	6	7	8	9	10
x_i	0	2	1	6	7	3	8	9	4	5
y_i	2	3	3	3	3	4	4	4	5	—
$g_i(x_i)$	0.101	0.713	0.605	0.743	0.212	0.781	0.035	0.003	0.327	1

There is even a clever way to replace the two simulations of U and I with a single simulation. Simulate Y with the uniform distribution on the interval [0, 1], and let I be the greatest integer less than or equal to $nY + 1$. Then let $U = nY + 1 - I$. (See Exercise 19.)

As an example, suppose that we simulate Y with the uniform distribution on the interval [0, 1], and we obtain $Y = 0.4694$. Then $I = 5$ and $U = 0.694$. Since $0.694 > g_5(x_5) = 0.212$, we set $X = y_5 = 3$. Figure 12.6 shows a histogram of 10,000 simulated values using the alias method. ◀

All of the overhead required to set up the alias method is worth the effort only if we are going to simulate many random variables with the same discrete distribution.

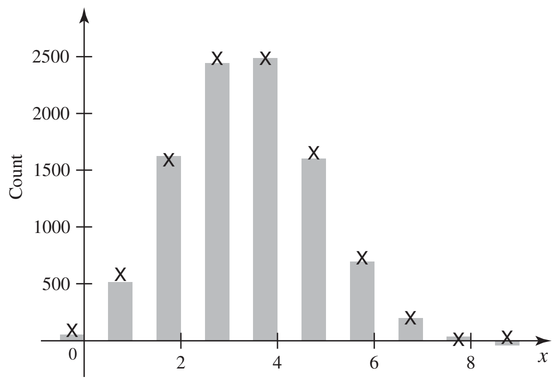


Figure 12.6 Histogram of 10,000 simulated binomial random variables in Example 12.3.13. The X marks appear at heights equal to $10{,}000 f(x)$ to illustrate the close agreement of the simulated and actual distributions.

Summary

We have seen several examples of how to transform pseudo-random uniform variables into pseudo-random variables with other distributions. The acceptance/rejection method is widely applicable, but it might require many rejected simulations for each accepted one. Also, we have seen how we can simulate random variables that are functions of other random variables (such as a noncentral t random variable). Several examples illustrated how we can make use of simulated random variables with some of the common distributions. Readers who desire a thorough treatment of the generation of pseudo-random variables with distributions other than uniform can consult Devroye (1986).

Exercises

1. Return to Exercise 10 in Sec. 12.2. Now that we know how to simulate exponential random variables, perform the simulation developed in that exercise as follows:

- **a.** Perform $v_0 = 2000$ simulations and compute both the estimate of θ and its simulation standard error.
- **b.** Suppose that we want our estimator of θ to be within 0.01 of θ with probability 0.99. How many simulations should we perform?

2. Describe how to convert a random sample $U_1, \dots, U_n$ from the uniform distribution on the interval [0, 1] to a random sample of size n from the uniform distribution on the interval $[a, b]$.

3. Show how to use the probability integral transformation to simulate random variables with the two p.d.f.'s in Eq. (12.3.3).

4. Show how to simulate Cauchy random variables using the probability integral transformation.

5. Prove that the expected number of iterations of the acceptance/rejection method until the first acceptance is k. (*Hint:* Think of each iteration as a Bernoulli trial. What is the expected number of trials (not failures) until the first success?)

6. a. Show how to simulate a random variable having the Laplace distribution with parameters 0 and 1. The p.d.f. of the Laplace distribution with parameters θ and σ is given in Eq. (10.7.5).

b. Show how to simulate a standard normal random variable by first simulating a Laplace random variable and then using acceptance/rejection. *Hint:* Maximize $e^{-x^2/2}/e^{-x}$ for $x \geq 0$, and notice that both distributions are symmetric around 0.

7. Suppose that you have available as many i.i.d. standard normal pseudo-random numbers as you desire. Describe how you could simulate a pseudo-random number with an F distribution with four and seven degrees of freedom.

8. Let X and Y be independent random variables with X having the t distribution with five degrees of freedom and Y having the t distribution with seven degrees of freedom. We are interested in $E(|X - Y|)$.

a. Simulate 1000 pairs of (X_i, Y_i) each with the above joint distribution and estimate $E(|X - Y|)$.

b. Use your 1000 simulated pairs to estimate the variance of $|X - Y|$ also.

c. Based on your estimated variance, how many simulations would you need to be 99 percent confident that your estimator of $E(|X - Y|)$ is within 0.01 of the actual mean?

9. Show how to use acceptance/rejection to simulate random variables with the following p.d.f.:

$$f(x) = \begin{cases} \frac{4}{3}x & \text{if } 0 < x \le 0.5, \\ \frac{2}{3} & \text{if } 0.5 < x \le 1.5, \\ \frac{8}{3} - \frac{4}{3}x & \text{if } 1.5 < x \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

10. Implement the simulation in Example 12.2.3 for the clinical trial of Example 2.1.4 on page 57. Simulate 5000 parameter vectors. Use a prior distribution with $\alpha_0 = 1$ and $\beta_0 = 1$. Estimate the probability that the imipramine group has the highest probability of no relapse. Calculate how many simulations you would need to be 95 percent confident that your estimator is within 0.01 of the true probability.

11. In Example 12.3.7, we simulated the τ values by first simulating gamma random variables with parameters $(n - p)/2$ and 1. Suppose that our statistical software allows us to simulate χ^2 random variables instead. Which χ^2 distribution should we use and how would we convert the simulated χ^2's to have the appropriate gamma distribution?

12. Use the blood pressure data in Table 9.2 that was described in Exercise 10 of Sec. 9.6. Suppose now that we are not confident that the variances are the same for the two treatment groups. Perform a simulation of the sort done in Example 12.3.8 to obtain a sample from the posterior distribution of the parameters when we allow the variances to be unequal.

a. Draw a plot of the sample c.d.f. of the absolute value of the difference between the two group means.

b. Draw a histogram of the logarithm of the ratio of the two variances to see how close together they seem to be.

13. Let F^{-1} be defined as in Eq. (12.3.7). Let U have the uniform distribution on the interval [0, 1]. Prove that $F^{-1}(U)$ has the c.d.f. in Eq. (12.3.6).

14. Refer to the three curves in Fig. 12.4. Call those three sample c.d.f.'s $G_{v,1}$, $G_{v,2}$, and $G_{v,3}$, and call the three c.d.f.'s that they estimate G_1, G_2, and G_3. Use the Glivenko-Cantelli lemma (Theorem 10.6.1) to show that

$$\Pr\left(|G_{v,i}(x) - G_i(x)| \le 0.0082, \text{ for all } x \text{ and all } i\right)$$

is about 0.9979 or larger. *Hint:* Use the Bonferroni inequality (Theorem 1.5.8).

15. Prove that the acceptance/rejection method works for discrete distributions. That is, let f and g be p.f.'s rather than p.d.f.'s, but let the rest of the acceptance/rejection method be exactly as stated. *Hint:* The proof can be translated by replacing integrals over x by sums. Integrals over u should be left as integrals.

16. Describe how to use the discrete version of the probability integral transformation to simulate a Poisson pseudo-random variable with mean θ.

17. Let f be a p.f., and assume that Eq. (12.3.8) holds, where each g_i is another p.f. Assume that X is simulated using the method described immediately after Eq. (12.3.8). Prove that X has the p.f. f.

18. Use the alias method to simulate a random variable having the Poisson distribution with mean 5. Use the table of Poisson probabilities in the back of the book, and assume that 16 is the largest value that a Poisson random variable can equal. Assume that all of the probability not accounted for by the values $0, \ldots, 15$ is the value of the p.f. at $k = 16$.

19. Let Y have the uniform distribution on the interval [0, 1]. Define I to be the greatest integer less than or equal to $nY + 1$, and define $U = nY + 1 - I$. Prove that I and U are independent and that U has uniform distribution on the interval [0, 1].

12.4 Importance Sampling

Many integrals can usefully be rewritten as means of functions of random variables. If we can simulate large numbers of random variables with the appropriate distributions, we can use these to estimate integrals that might not be possible to compute in closed form.

Simulation methods are particularly well suited to estimating means of random variables. If we can simulate many random variables with the appropriate distribution,

we can average the simulated values to estimate the mean. Because means of random variables with continuous distributions are integrals, we might wonder whether other integrals can also be estimated by simulation methods. In principle, all finite integrals can be estimated by simulation, although some care is needed to insure that the simulation results have finite variance.

Suppose that we wish to calculate $\int_a^b g(x)\,dx$ for some function g with a and b both finite. We can rewrite this integral as

$$\int_a^b g(x)\,dx = \int_a^b (b-a)g(x)\frac{1}{b-a}\,dx = E[(b-a)g(X)], \tag{12.4.1}$$

where X is a random variable with the uniform distribution on the interval $[a, b]$. A simple Monte Carlo method is to simulate a large number of pseudo-random values $X_1, \ldots, X_v$ with the uniform distribution on the interval $[a, b]$ and estimate the integral by $\frac{b-a}{v}\sum_{i=1}^{v} g(X_i)$. The method just described has two commonly recognized drawbacks. First, it cannot be applied to estimate integrals over unbounded regions. Second, it can be very inefficient. If g is much larger over one portion of the interval than over another, then the values $g(X_i)$ will have large variance, and it will take a very large value v to get a good estimator of the integral.

A method that attempts to overcome both of the shortcomings just mentioned is called *importance sampling*. The idea of importance sampling is to do something very much like what we did in Eq. (12.4.1). That is, we shall rewrite the integral as the mean of some function of X, where X has a distribution that we can simulate easily.

Suppose that we are able to simulate a pseudo-random variable X with the p.d.f. f where $f(x) > 0$ whenever $g(x) > 0$. Then we can write

$$\int g(x)\,dx = \int \frac{g(x)}{f(x)} f(x)\,dx = E(Y), \tag{12.4.2}$$

where $Y = g(X)/f(X)$. (If $f(x) = 0$ for some x such that $g(x) > 0$, then the two integrals in Eq. (12.4.2) might not be equal.) If we simulate v independent values $X_1, \ldots, X_v$ with the p.d.f. f, we can estimate the integral by $\frac{1}{v}\sum_{i=1}^{v} Y_i$ where $Y_i = g(X_i)/f(X_i)$. The p.d.f. f is called the *importance function*. It is acceptable, although inefficient, to have $f(x) > 0$ for some x such that $g(x) = 0$. The key to efficient importance sampling is choosing a good importance function. The smaller the variance of Y, the better the estimator should be. That is, we would like $g(X)/f(X)$ to be close to being a constant random variable.

Example 12.4.1 Choosing an Importance Function. Suppose that we want to estimate $\int_0^1 e^{-x}/\,(1+x^2)dx$. Here are five possible choices of importance function:

$$\begin{aligned}
f_0(x) &= 1, \quad \text{for } 0 < x < 1,\\
f_1(x) &= e^{-x}, \quad \text{for } 0 < x < \infty,\\
f_2(x) &= (1+x^2)^{-1}/\pi, \quad \text{for } -\infty < x < \infty,\\
f_3(x) &= e^{-x}/(1-e^{-1}), \quad \text{for } 0 < x < 1,\\
f_4(x) &= 4(1+x^2)^{-1}/\pi, \quad \text{for } 0 < x < 1.
\end{aligned}$$

Each of these p.d.f.'s is positive wherever g is positive, and each one can be simulated using the probability integral transformation. As an example, we have simulated 10,000 uniforms on the interval [0, 1], $U^{(1)}, \ldots, U^{(10{,}000)}$. We then applied the five probability integral transformations to this single set of uniforms so that our comparisons do not suffer from variation due to different underlying uniform samples.

Table 12.5 Monte Carlo estimates and $\hat{\sigma}_j$ for Example 12.4.1

j	0	1	2	3	4
$\overline{Y}_j$	0.5185	0.5110	0.5128	0.5224	0.5211
$\hat{\sigma}_j$	0.2440	0.4217	0.9312	0.0973	0.1409

Since the five p.d.f.'s are positive over different ranges, we should define

$$g(x) = \begin{cases} e^{-x}/(1+x^2) & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let F_j stand for the c.d.f. corresponding to f_j, and let $X_j^{(i)} = F_j^{-1}(U^{(i)})$ for $i = 1, \ldots, 10{,}000$ and $j = 0, \ldots, 4$. Let $Y_j^{(i)} = g(X_j^{(i)})/f_j(X_j^{(i)})$. Then we obtain five different estimators of $\int g(x)\,dx$, namely,

$$\overline{Y}_j = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} Y_j^{(i)}, \text{ for } j = 0, \ldots, 4.$$

For each j, we also compute the sample variance of the $Y_j^{(i)}$ values,

$$\hat{\sigma}_j^2 = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} (Y_j^{(i)} - \overline{Y}_j)^2.$$

The simulation standard error of $\overline{Y}_j$ is $\hat{\sigma}_j/100$. We list the five estimates together with the corresponding values of $\hat{\sigma}_j$ in Table 12.5. The estimates are relatively close together, but some values of $\hat{\sigma}_j$ are almost 10 times others. This can be understood in terms of how well each f_j approximates the function g. First, note that the two worst cases are those in which f_j is positive on an unbounded interval. This causes us to simulate a large number of $X_j^{(i)}$ values for which $g(X_j^{(i)}) = 0$ and hence $Y_j^{(i)} = 0$. This is highly inefficient. For example, with $j = 2$, 75 percent of the $X_2^{(i)}$ values are outside of the interval (0, 1). The remaining $Y_2^{(i)}$ values must be very large in order for the average to come out near the correct answer. In other words, because the $Y_2^{(i)}$ values are so spread out (they range between 0 and π), we get a large value of $\hat{\sigma}_2$. On the other hand, with $j = 3$, there are no 0 values for $Y_3^{(i)}$. Indeed, the $Y_3^{(i)}$ values only range from 0.3161 to 0.6321. This allows $\hat{\sigma}_3$ to be quite small. The goal in choosing an importance function is to make the $Y^{(i)}$ values have small variance. This is achieved by making the ratio g/f as close to constant as we can. ◀

Example 12.4.2 **Calculating a Mean with No Closed-Form Expression.** Let X have the gamma distribution with parameters α and 1. Suppose that we want the mean of $1/(1 + X + X^2)$. We might wish to think of this mean as

$$\int_0^\infty \frac{1}{1 + x + x^2} f_\alpha(x)\,dx, \qquad (12.4.3)$$

where f_α is the p.d.f. of the gamma distribution with parameters α and 1. If α is not small, $f_\alpha(x)$ is close to 0 near $x = 0$ and is only sizeable for x near α. For large x,

$1/(1+x+x^2)$ is a lot like $1/x^2$. If α and x are both large, the integrand in (12.4.3) is approximately $x^{-2}f_\alpha(x)$. Since $x^{-2}f_\alpha(x)$ is a constant times $f_{\alpha-2}(x)$, we could do importance sampling with importance function $f_{\alpha-2}$. For example, with $\alpha = 5$, we simulate 10,000 pseudo-random variables $X^{(1)}, \ldots, X^{(10,000)}$ having the gamma distribution with parameters 3 and 1. The sample mean of $[1/(1+X^{(i)}+X^{(i)2})]f_5(X^{(i)})/f_3(X^{(i)})$ is 0.05184 with sample standard deviation 0.01465. For comparison, we also simulate 10,000 pseudo-random variables $Y^{(1)}, \ldots, Y^{(10,000)}$ with the gamma distribution having parameters 5 and 1. The average of the values of $1/(1+Y^{(i)}+Y^{(i)2})$ is 0.05226 with sample standard deviation 0.05103, about 3.5 times as large as we get using the f_3 importance function. With $\alpha = 3$, however, the two methods have nearly equal sample standard deviations. With $\alpha = 10$, the importance sampling has sample standard deviation about one-tenth as large as sampling directly from the distribution of X. As we noted earlier, when α is large, $1/x^2$ is a better approximation to $1/(1+x+x^2)$ than it is when α is small. ◀

Example 12.4.3 **Bivariate Normal Probabilities.** Let (X_1, X_2) have a bivariate normal distribution, and suppose that we are interested in the probability of the event $\{X_1 \le c_1, X_2 \le c_2\}$ for specific values c_1, c_2. In general, we cannot explicitly calculate the double integral

$$\int_{-\infty}^{c_2}\int_{-\infty}^{c_1} f(x_1, x_2)\,dx_1\,dx_2, \tag{12.4.4}$$

where $f(x_1, x_2)$ is the joint p.d.f. of (X_1, X_2). We can write the joint p.d.f. as $f(x_1, x_2) = g_1(x_1|x_2)f_2(x_2)$, where g_1 is the conditional p.d.f. of X_1 given $X_2 = x_2$ and f_2 is the marginal p.d.f. of X_2. Both of these p.d.f.'s are normal p.d.f.'s, as we learned in Sec. 5.10. In particular, the conditional distribution of X_1 given $X_2 = x_2$ is the normal distribution with mean and variance given by Eq. (5.10.8). We can explicitly perform the inner integration in (12.4.4) as

$$\begin{aligned}\int_{-\infty}^{c_1} f(x_1, x_2)\,dx_1 &= \int_{-\infty}^{c_1} g(x_1|x_2)f_2(x_2)\,dx_1\\ &= f_2(x_2)\Phi\left(\frac{c_1-\mu_1-\rho\sigma_1(x_2-\mu_2)/\sigma_2}{\sigma_1(1-\rho)^{1/2}}\right),\end{aligned}$$

where Φ is the standard normal c.d.f. The integral in (12.4.4) is then the integral of this last expression with respect to x_2. An efficient importance function might be the conditional p.d.f. of X_2 given that $X_2 \le c_2$. That is, let h be the p.d.f.

$$h(x_2) = \frac{(2\pi\sigma_2^2)^{-1/2}\exp\left(-\dfrac{1}{2\sigma_2^2}(x_2-\mu_2)^2\right)}{\Phi\left(\dfrac{c_2-\mu_2}{\sigma_2}\right)}, \quad \text{for } -\infty < x_2 \le c_2. \tag{12.4.5}$$

It is not difficult to see that if U has the uniform distribution on the interval [0, 1], then

$$W = \mu_2 + \sigma_2\Phi^{-1}\left[U\Phi\left(\frac{c_2-\mu_2}{\sigma_2}\right)\right] \tag{12.4.6}$$

has the p.d.f. h. (See Exercise 5.) If we use h as an importance function and simulate $W^{(1)}, \ldots, W^{(v)}$ with this p.d.f., then our estimator of the integral (12.4.4) is

$$\frac{1}{v}\sum_{i=1}^{v}\Phi\left(\frac{c_1-\mu_1-\rho\sigma_1(W^{(i)}-\mu_2)/\sigma_2}{\sigma_1(1-\rho)^{1/2}}\right)\Phi\left(\frac{c_2-\mu_2}{\sigma_2}\right). \quad ◀$$

It is not always possible to guarantee that an importance sampling estimator will have finite variance. In the examples in this section, we have managed to find importance functions with the following property. The ratio of the function being integrated to the importance function is bounded. This property guarantees finite variance for the importance sampling estimator. (See Exercise 8.)

Stratified Importance Sampling

Suppose that we are trying to estimate $\theta = \int g(x)\,dx$, and that we contemplate using the importance function f. The simulation variance of the importance sampling estimator of θ arises from the variance of $Y = g(X)/f(X)$, where X has the p.d.f. f. Indeed, if we simulate an importance sample of size n, the simulation variance of our estimator is σ^2/n, where $\sigma^2 = \text{Var}(Y)$. Stratified importance sampling attempts to reduce the simulation variance by splitting θ into $\theta = \sum_{j=1}^{k} \theta_j$ and then estimating each θ_j with much smaller simulation variance.

The stratified importance sampling algorithm is easiest to describe when X is simulated using the probability integral transformation. Let F be the c.d.f. corresponding to the p.d.f. f. First, we split θ as follows. Define $q_0 = -\infty$, $q_j = F^{-1}(j/k)$ for $j = 1, \ldots, k-1$, and $q_k = \infty$. Then define

$$\theta_j = \int_{q_{j-1}}^{q_j} g(x)\,dx,$$

for $j = 1, \ldots, k$. Clearly, $\theta = \sum_{j=1}^{k} \theta_j$. Next, we estimate each θ_j by importance sampling using the same importance function f, but restricted to the range of integration for θ_j. That is, we estimate θ_j using importance sampling with the importance function

$$f_j(x) = \begin{cases} kf(x) & \text{if } q_{j-1} \le x < q_j, \\ 0 & \text{otherwise.} \end{cases}$$

(See Exercise 9 to see that f_j is indeed a p.d.f.) To simulate a random variable with the p.d.f. f_j, let V have the uniform distribution on the interval $[(j-1)/k, j/k]$ and set $X_j = F^{-1}(V)$. The reader can prove (see Exercise 9) that X_j has the p.d.f. f_j. Let σ_j^2 be the variance of $g(X_j)/f_j(X_j)$. Suppose that, for each $j = 1, \ldots, k$, we simulate an importance sample of size m with the same distribution as X_j. The variance of the estimator of θ_j will be σ_j^2/m. Since the k estimators of $\theta_1, \ldots, \theta_k$ are independent, the variance of the estimator of θ will be $\sum_{j=1}^{k} \sigma_j^2/m$. To facilitate comparison to nonstratified importance sampling, let $n = mk$. Stratification will be an improvement if its variance is smaller than σ^2/n. Since $n = mk$, we would like to prove that at least

$$\sigma^2 \ge k \sum_{j=1}^{k} \sigma_j^2, \tag{12.4.7}$$

and preferably with strict inequality.

To prove (12.4.7), we note a close connection between the random variables X_j with the p.d.f. f_j and X with the p.d.f. f. Let J be a random variable with the discrete uniform distribution on the integers $1, \ldots, k$. Define $X^* = X_J$, so that the conditional p.d.f. of X^* given $J = j$ is f_j. You can prove (Exercise 11) that X^* and X have the same p.d.f. Let $Y = g(X)/f(X)$ and

$$Y^* = \frac{g(X^*)}{f_J(X^*)} = \frac{g(X^*)}{kf(X^*)}.$$

Then $\text{Var}(Y^*|J = j) = \sigma_j^2$ and kY^* has the same distribution as Y. So,

$$\sigma^2 = \text{Var}(Y) = \text{Var}(kY^*) = k^2\,\text{Var}(Y^*). \tag{12.4.8}$$

Theorem 4.7.4 says that

$$\text{Var}(Y^*) = E\,\text{Var}(Y^*|J) + \text{Var}[E(Y^*|J)]. \tag{12.4.9}$$

By construction, $E(Y^*|J = j) = \theta_j$ and $\text{Var}(Y^*|J = j) = \sigma_j^2$. Also, $\text{Var}[E(Y^*|J)] \ge 0$ with strict inequality if the θ_j are not all the same. Since $\Pr(J = j) = 1/k$ for $j = 1, \ldots, k$, we have

$$E\,\text{Var}(Y^*|J) = \frac{1}{k}\sum_{j=1}^{k}\sigma_j^2. \tag{12.4.10}$$

Combining Eqs. (12.4.8), (12.4.9), and (12.4.10), we obtain (12.4.7), with strict inequality if the θ_j are not all equal.

Example 12.4.4 Illustration of Stratified Importance Sampling. Consider the integral that we wanted to estimate in Example 12.4.1. The best importance function appeared to be f_3, with a simulation standard error of $\hat{\sigma}_3/100 = 9.73 \times 10^{-4}$. In the present example, we allocate 10,000 simulations among $k = 10$ subsets of size $m = 1000$ each and do stratified importance sampling by dividing the range of integration [0, 1] into 10 equal-length subintervals. Doing this, we get a Monte Carlo estimate of the integral of 0.5248. To estimate the simulation standard error, we need to estimate each σ_j by $\hat{\sigma}_j^*$ and compute $\sum_{j=1}^{10}\hat{\sigma}_j^{*2}/1000$. In the simulation that we are discussing, the simulation standard error for stratified importance sampling is 1.05×10^{-4}, about one-tenth as small as the unstratified version. We can also do stratified importance sampling using $k = 100$ subsets of size $m = 100$. In our simulation, the estimate of the integral is the same with simulation standard error of 1.036×10^{-5}. ◀

The reason that stratified importance sampling works so well in Example 12.4.4 is that the function $g(x)/f_3(x)$ is monotone, and this makes θ_j change about as much as it can as j changes. Hence, $\text{Var}[E(Y^*|J)]$ is large, making stratification very effective.

Summary

We introduced the method of importance sampling for calculating integrals by simulation. The idea of importance sampling for estimating $\int g(x)\,dx$ is to choose a p.d.f. f from which we can simulate and such that $g(x)/f(x)$ is nearly constant. Then we rewrite the integral as $\int [g(x)/f(x)]f(x)\,dx$. We can estimate this last integral by averaging $g(X^{(i)})/f(X^{(i)})$ where $X^{(1)}, \ldots, X^{(v)}$ form a random sample with the p.d.f. f. A stratified version of importance sampling can produce estimators with even smaller variance.

Exercises

1. Prove that the formula in Eq. (12.4.1) is the same as importance sampling in which the importance function is the p.d.f. of the uniform distribution on the interval $[a, b]$.

2. Let g be a function, and suppose that we wish to compute the mean of $g(X)$ where X has the p.d.f. f. Suppose that we can simulate pseudo-random values with the p.d.f. f. Prove that the following are the same:

- Simulate $X^{(i)}$ values with the p.d.f. f, and average the values of $g(X^{(i)})$ to obtain the estimator.
- Do importance sampling with importance function f to estimate the integral $\int g(x)f(x)\,dx$.

3. Let Y have the F distribution with m and n degrees of freedom. We wish to estimate $\Pr(Y > c)$. Consider the p.d.f.

$$f(x) = \begin{cases} \frac{(n/2)c^{n/2}}{x^{n/2+1}} & \text{if } x > c, \\ 0 & \text{otherwise.} \end{cases}$$

a. Explain how to simulate pseudo-random numbers with the p.d.f. f.

b. Explain how to estimate $\Pr(Y > c)$ using importance sampling with the importance function f.

c. Look at the form of the p.d.f. of Y, Eq. (9.7.2), and explain why importance sampling might be more efficient than sampling i.i.d. F random variables with m and n degrees of freedom if c is not small.

4. We would like to calculate the integral $\int_0^\infty \log(1 + x)\exp(-x)\, dx$.

a. Simulate 10,000 exponential random variables with parameter 1 and use these to estimate the integral. Also, find the simulation standard error of your estimator.

b. Simulate 10,000 gamma random variables with parameters 1.5 and 1 and use these to estimate the integral (importance sampling). Find the simulation standard error of the estimator. (In case you do not have the gamma function available, $\Gamma(1.5) = \sqrt{\pi}/2$.)

c. Which of the two methods appears to be more efficient? Can you explain why?

5. Let U have the uniform distribution on the interval [0, 1]. Show that the random variable W defined in Eq. (12.4.6) has the p.d.f. h defined in Eq. (12.4.5).

6. Suppose that we wish to estimate the integral

$$\int_1^\infty \frac{x^2}{\sqrt{2\pi}} e^{-0.5x^2}\, dx.$$

In parts (a) and (b) below, use simulation sizes of 1000.

a. Estimate the integral by importance sampling using random variables having a truncated normal distribution. That is, the importance function is

$$\frac{1}{\sqrt{2\pi}[1 - \Phi(1)]} e^{-0.5x^2}, \quad \text{for } x > 1.$$

b. Estimate the integral by importance sampling using random variables with the p.d.f. $x\exp(0.5[1 - x^2])$, for $x > 1$. *Hint:* Prove that such random variables can be obtained as follows: Start with a random variable that has the exponential distribution with parameter 0.5, add 1, then take the square root.

c. Compute and compare simulation standard errors for the two estimators in parts (a) and (b). Can you explain why one is so much smaller than the other?

7. Let (X_1, X_2) have the bivariate normal distribution with both means equal to 0, both variances equal to 1, and the correlation equal to 0.5. We wish to estimate $\theta = \Pr(X_1 \le 2, X_2 \le 1)$ using simulation.

a. Simulate a sample of 10,000 bivariate normal vectors with the above distribution. Use the proportion of vectors satisfying the two inequalities $X_1 \le 2$ and $X_2 \le 1$ as the estimator Z of θ. Also compute the simulation standard error of Z.

b. Use the method described in Example 12.4.3 with 10,000 simulations to produce an alternative estimator Z' of θ. Compute the simulation standard error of Z' and compare Z' to the estimate in part (a).

8. Suppose that we wish to approximate the integral $\int g(x)\, dx$. Suppose that we have a p.d.f. f that we shall use as an importance function. Suppose that $g(x)/f(x)$ is bounded. Prove that the importance sampling estimator has finite variance.

9. Let F be a continuous strictly increasing c.d.f. with p.d.f. f. Let V have the uniform distribution on the interval $[a, b]$ with $0 \le a < b \le 1$. Prove that the p.d.f. of $X = F^{-1}(V)$ is $f(x)/(b - a)$ for $F^{-1}(a) \le x \le F^{-1}(b)$. (If $a = 0$, let $F^{-1}(a) = -\infty$. If $b = 1$, let $F^{-1}(b) = \infty$.)

10. For the situation described in Exercise 6, use stratified importance sampling as follows: Divide the interval $(1, \infty)$ into five intervals that each have probability 0.2 under the importance distribution. Sample 200 observations from each interval. Compute the simulation standard error. Compare this simulation to the simulation in Exercise 6 for each of parts (a) and (b).

11. In the notation used to develop stratified importance sampling, prove that $X^* = X_J$ and X have the same distribution. *Hint:* The conditional p.d.f. of X^* given $J = j$ is f_j. Use the law of total probability.

12. Consider again the situation described in Exercise 15 of Sec. 12.2. Suppose that W_u has the Laplace distribution with parameters $\theta = 0$ and $\sigma = 0.1u^{1/2}$. See Eq. (10.7.5) for the p.d.f.

a. Prove that the m.g.f. of W_u is

$$\psi(t) = \left(1 - \frac{t^2 u}{100}\right)^{-1}, \quad \text{for } -10u^{-1/2} < t < 10u^{-1/2}.$$

b. Let $r = 0.06$ be the risk-free interest rate. Simulate a large number v of values of W_u with $u = 1$ and use these to estimate the price of an option to buy one share of this stock at time $u = 1$ in the future for the current price S_0. Also compute the simulation standard error.

c. Use importance sampling to improve on the simulation in part (b). Instead of simulating W_u values directly, simulate from the conditional distribution of W_u given that $S_u > S_0$. How much smaller is the simulation standard error?

13. The method of *control variates* is a technique for reducing the variance of a simulation estimator. Suppose that we wish to estimate $\theta = E(W)$. A control variate is another random variable V that is positively correlated with W and whose mean μ we know. Then, for every constant $k > 0$, $E(W - kV + k\mu) = \theta$. Also, if k is chosen carefully, $\mathrm{Var}(W - kV + k\mu) < \mathrm{Var}(W)$. In this exercise, we shall see how to use control variates for importance sampling, but the method is very general. Suppose that we wish to compute $\int g(x)\,dx$, and we wish to use the importance function f. Suppose that there is a function h such that h is similar to g but $\int h(x)\,dx$ is known to equal the value c. Let k be a constant. Simulate $X^{(1)}, \ldots, X^{(v)}$ with the p.d.f. f, and define

$$W^{(i)} = \frac{g(X^{(i)})}{f(X^{(i)})},$$
$$V^{(i)} = \frac{h(X^{(i)})}{f(X^{(i)})},$$
$$Y^{(i)} = W^{(i)} - kV^{(i)},$$

for all i. Our estimator of $\int g(x)\,dx$ is then $Z = \frac{1}{v}\sum_{i=1}^{v} Y^{(i)} + kc$.

a. Prove that $E(Z) = \int g(x)\,dx$.

b. Let $\mathrm{Var}(W^{(i)}) = \sigma_W^2$ and $\mathrm{Var}(V^{(i)}) = \sigma_V^2$. Let ρ be the correlation between $W^{(i)}$ and $V^{(i)}$. Prove that the value of k that makes $\mathrm{Var}(Z)$ the smallest is $k = \sigma_W \rho / \sigma_V$.

14. Suppose that we wish to integrate the same function $g(x)$ as in Example 12.4.1.

a. Use the method of control variates that was described in Exercise 13 to estimate $\int g(x)\,dx$. Let $h(x) = 1/(1 + x^2)$ for $0 < x < 1$, and $k = e^{-0.5}$. (This makes h about the same size as g.) Let $f(x)$ be the function f_3 in Example 12.4.1. How does the simulation standard error using control variates compare to not using control variates?

b. Estimate the variances and correlation of the $W^{(i)}$'s and $V^{(i)}$'s (notation of Exercise 13) to see what a good value for k might be.

15. The method of *antithetic variates* is a technique for reducing the variance of simulation estimators. Antithetic variates are negatively correlated random variables that share a common mean and common variance. The variance of the average of two antithetic variates is smaller than the variance of the average of two i.i.d. variables. In this exercise, we shall see how to use antithetic variates for importance sampling, but the method is very general. Suppose that we wish to compute $\int g(x)\,dx$, and we wish to use the importance function f. Suppose that we generate pseudo-random variables with the p.d.f. f using the probability integral transformation. That is, for $i = 1, \ldots, v$, let $X^{(i)} = F^{-1}(U^{(i)})$, where $U^{(i)}$ has the uniform distribution on the interval [0, 1] and F is the c.d.f. corresponding to the p.d.f. f. For each $i = 1, \ldots, v$, define

$$T^{(i)} = F^{-1}(1 - U^{(i)}),$$
$$W^{(i)} = \frac{g(X^{(i)})}{f(X^{(i)})},$$
$$V^{(i)} = \frac{g(T^{(i)})}{f(T^{(i)})},$$
$$Y^{(i)} = 0.5\left[W^{(i)} + V^{(i)}\right].$$

Our estimator of $\int g(x)\,dx$ is then $Z = \frac{1}{v}\sum_{i=1}^{v} Y^{(i)}$.

a. Prove that $T^{(i)}$ has the same distribution as $X^{(i)}$.

b. Prove that $E(Z) = \int g(x)\,dx$.

c. If $g(x)/f(x)$ is a monotone function, explain why we would expect $W^{(i)}$ and $V^{(i)}$ to be negatively correlated.

d. If $W^{(i)}$ and $V^{(i)}$ are negatively correlated, show that $\mathrm{Var}(Z)$ is less than the variance one would get with $2v$ simulations without antithetic variates.

16. Use the method of antithetic variates that was described in Exercise 15. Let $g(x)$ be the function that we tried to integrate in Example 12.4.1. Let $f(x)$ be the function f_3 in Example 12.4.1. Estimate $\mathrm{Var}(Y^{(i)})$, and compare it to $\hat{\sigma}_3^2$ from Example 12.4.1.

17. For each of the exercises in this section that requires a simulation, see if you can think of a way to use control variates or antithetic variates to reduce the variance of the simulation estimator.

★ 12.5 Markov Chain Monte Carlo

The techniques described in Sec. 12.3 for generating pseudo-random numbers with particular distributions are most useful for univariate distributions. They can be applied in many multivariate cases, but they often become unwieldy. A method based on Markov chains (see Sec. 3.10) became popular after publications by Metropolis et al. (1953) and Gelfand and Smith (1990). We shall present only the simplest form of Markov chain Monte Carlo in this section.

The Gibbs Sampling Algorithm

We shall begin with an attempt to simulate a bivariate distribution. Suppose that the joint p.d.f. of (X_1, X_2) is $f(x_1, x_2) = cg(x_1, x_2)$, where we know the function g but not necessarily the value of the constant c. This type of situation arises often when computing posterior distributions. If X_1 and X_2 are the parameters, the function g might be the product of the prior p.d.f. times the likelihood function (in which the data are treated as known values). The constant $c = 1/\int g(x_1, x_2)\, dx_1\, dx_2$ makes $cg(x_1, x_2)$ the posterior p.d.f. Often it is difficult to compute c, although the methods of Sec. 12.4 might be helpful. Even if we can approximate the constant c, there are other features of the posterior distribution that we might not be able to compute easily, so simulation would be useful.

If the function $g(x_1, x_2)$ has a special form, then there is a powerful algorithm for simulating vectors with the p.d.f. f. The required form can be described as follows: First, consider $g(x_1, x_2)$ as a function of x_1 for fixed x_2. This function needs to look like a p.d.f. (for X_1) from which we know how to simulate pseudo-random values. Similarly, if we consider $g(x_1, x_2)$ as a function of x_2 for fixed x_1, the function needs to look like a p.d.f. for X_2 from which can simulate.

Example 12.5.1 Sample from a Normal Distribution. Suppose that we have observed a sample from the normal distribution with unknown mean μ and unknown precision τ. Suppose that we use a natural conjugate prior of the form described in Sec. 8.6. The product of the prior and the likelihood is given by Eq. (8.6.7) without the appropriate constant factor. We reproduce a version of that equation here for convenience:

$$\xi(\mu, \tau|\boldsymbol{x}) \propto \tau^{\alpha_1+1/2-1} \exp\left(-\tau\left[\frac{1}{2}\lambda_1(\mu-\mu_1)^2+\beta_1\right]\right),$$

where α_1, β_1, μ_1, and λ_1 are known values once the data have been observed. Considering this as a function of μ for fixed τ, it looks like the p.d.f. of the normal distribution with mean μ_1 and variance $(\tau\lambda_1)^{-1}$. Considering it as a function of τ for fixed μ, it looks like the p.d.f. of the gamma distribution with parameters $\alpha_1 + 1/2$ and $\lambda_1(\mu-\mu_1)^2/2 + \beta_1$. Both of these distributions are easy to simulate. ◀

When we consider $g(x_1, x_2)$ as a function of x_1 for fixed x_2, we are looking at the conditional p.d.f. of X_1 given $X_2 = x_2$, except for a multiplicative factor that does not depend on x_1. (See Exercise 1.) Similarly, when we consider $g(x_1, x_2)$ as a function of x_2 for fixed x_1, we are looking at the conditional p.d.f. of X_2 given $X_1 = x_1$.

Once we have determined that the function $g(x_1, x_2)$ has the desired form, our algorithm proceeds as follows:

1. Pick a starting value $x_2^{(0)}$ for X_2, and set $i = 0$.
2. Simulate a new value $x_1^{(i+1)}$ from the conditional distribution of X_1 given $X_2 = x_2^{(i)}$.
3. Simulate a new value $x_2^{(i+1)}$ from the conditional distribution of X_2 given $X_1 = x_1^{(i+1)}$.
4. Replace i by $i + 1$ and return to step 2.

The algorithm typically terminates when i reaches a sufficiently large value. Although there currently are no truly satisfactory convergence criteria, we shall introduce one convergence criterion later in this section. This algorithm is commonly called *Gibbs*

sampling. The name derives from an early use of the technique by Geman and Geman (1984) for sampling from a distribution that was known as the Gibbs distribution.

Some Theoretical Justification

So far, we have given no justification for the Gibbs sampling algorithm. The justification stems from the fact that the successive pairs $(x_1^{(1)}, x_2^{(1)})$, $(x_1^{(2)}, x_2^{(2)}) \dots$ form the observed sequence of states from a Markov chain. This Markov chain is much more complicated than any of the Markov chains encountered in Sec. 3.10 for two reasons. First, the states are two-dimensional, and second, the number of possible states is infinite rather than finite. Even so, one can easily recognize the basic structure of a Markov chain in the description of the Gibbs sampling algorithm. Suppose that i is the current value of the iteration index. The conditional distribution of the next state pair $(X_1^{(i+1)}, X_2^{(i+1)})$ given all of the available state pairs $(X_1^{(1)}, X_2^{(1)}), \dots, (X_1^{(i)}, X_2^{(i)})$ depends only on $(X_1^{(i)}, X_2^{(i)})$, the current state pair. This is the same as the defining property of finite Markov chains in Sec. 3.10.

Even if we agree that the sequence of pairs forms a Markov chain, why should we believe that they come from the desired distribution? The answer lies in a generalization of the second part of Theorem 3.10.4 to more general Markov chains. The generalization is mathematically too involved to present here, and it requires conditions that involve concepts that we have not introduced in this book.

Nevertheless, the Gibbs sampler is constructed from a joint distribution that one can show (see Exercise 2) is a stationary distribution for the resulting Markov chain. For the cases that we illustrate in this book, the distribution of the Gibbs sampler Markov chain does indeed converge to this stationary distribution as the number of transitions increases. (For a more general discussion, see Tierney, 1994.) Because of the close connection with Markov chains, Gibbs sampling (and several related techniques) are often called *Markov chain Monte Carlo*.

When Does the Markov Chain Converge?

Although the distribution of a Markov chain may converge to its stationary distribution, after any finite time the distribution will not necessarily be the stationary distribution. In general, the distribution will get pretty close to the stationary distribution in finite time, but how do we tell, in a particular application, if we have sampled the Markov chain long enough to be confident that we are sampling from something close to the stationary distribution? Much work has been done to address this question, but there is no foolproof method. Several methods for assessing convergence of a Markov chain in a Monte Carlo analysis were reviewed by Cowles and Carlin (1996). Here we present one simple technique.

Begin by sampling several versions of the Markov chain starting at k different initial values $x_{2,1}^{(0)}, \dots, x_{2,k}^{(0)}$. These k Markov chains will be useful not only for assessing convergence but also for estimating the variances of our simulation estimators. It is wise to choose the initial values $x_{2,1}^{(0)}, \dots, x_{2,k}^{(0)}$ to be quite spread out. This will help us to determine whether we have a Markov chain that is very slow to converge. Next, apply the Gibbs sampling algorithm starting at each of the k initial values. This gives us k independent Markov chains, all with the same stationary distribution. If the k Markov chains have been sampled for m iterations, we can think of the observed values of X_1 (or of X_2) as k samples of size m each. For ease of notation, let $T_{i,j}$ stand for either the value of X_1 or the value of X_2 from the jth iteration of the

ith Markov chain. (We shall repeat the following analysis once for X_1 and once for X_2.) Now, treat $T_{i,j}$ for $j = 1, \ldots, m$ as a sample of size m from the ith of k distributions for $i = 1, \ldots, k$. If we have sampled long enough for the Markov chains to have converged approximately, then all k of these distributions should be nearly the same. This suggests that we use the F statistic from the discussion of analysis of variance (Sec. 11.6) to measure how close the k distributions are. The F statistic can be written as $F = B/W$ where

$$B = \frac{m}{k-1}\sum_{i=1}^{k}(\overline{T}_{i+} - \overline{T}_{++})^2,$$

$$W = \frac{1}{k(m-1)}\sum_{i=1}^{k}\sum_{j=1}^{m}(T_{ij} - \overline{T}_{i+})^2.$$

Here we have used the same notation as in Sec. 11.6 in which the + subscript appears in a position wherever we have averaged over all values of the subscript in that position. If the k distributions are different, then F should be large. If the distributions are the same, then F should be close to 1. As we mentioned earlier, we compute two F statistics, one using the X_1 coordinates and one using the X_2 coordinates. Then we could declare that we have sampled long enough when both F statistics are simultaneously less than some number slightly larger than 1. Gelman et al. (1995) describe essentially this same procedure and recommend comparing the maximum of the two F statistics to $1 + 0.44m$. It is probably a good idea to start with at least $m = 100$ (if the iterations are fast enough) before beginning to compute the F statistics. This will help to avoid accidentally declaring success due to some "lucky" early simulations. The initial sequence of iterations of the Markov chain, before we declare convergence, is commonly called *burn-in*. After the burn-in iterations, one would typically treat the ensuing iterations as observations from the stationary distribution. It is common to discard the burn-in iterations because we are not confident that their distribution is close to the stationary distribution. Iterations of a Markov Chain are dependent, however, so one should not treat them as an i.i.d. sample. Even though we computed an F statistic from the various dependent observations, we did not claim that the statistic had an F distribution. Nor did we compare the statistic to a quantile of an F distribution to make our decision about convergence. We merely used the statistic as an ad hoc measure of how different the k Markov chains are.

Example 12.5.2

Nursing Homes in New Mexico. We shall use the data from Sec. 8.6 on the numbers of medical in-patient days in 18 nonrural nursing homes in New Mexico in 1988. There, we modeled the observations as a random sample from the normal distribution with unknown mean μ and unknown precision τ. We used a natural conjugate prior and found the posterior hyperparameters to be $\alpha_1 = 11$, $\beta_1 = 50925.37$, $\mu_1 = 183.95$, and $\lambda_1 = 20$. We shall illustrate the above convergence diagnostic for the Gibbs sampling algorithm described in Example 12.5.1. As we found in Example 12.5.1, the conditional distribution of μ given τ is the normal distribution with mean 183.95 and variance $(20\tau)^{-1}$. The conditional distribution of τ given μ is the gamma distribution with parameters 11.5 and $50925.37 + 20(\mu - 183.95)^2$. We shall start with the following $k = 5$ initial values for μ: 182.17, 227, 272, 137, 82. These were chosen by making a crude approximation to the posterior standard deviation of μ, namely, $(\beta_1/[\lambda_1\alpha_1])^{1/2} \approx 15$, and then using the posterior mean together with values 3 and 6 posterior standard deviations above and below the posterior mean. We have to run the five Markov chains to the $m = 2$ iteration before we can compute the F statistics.

In our simulation, at $m = 2$, the larger of the two F statistics was already as low as 0.8862, and it stayed very close to 1 all the way to $m = 100$, at which time it seemed clear that we should stop the burn-in. ◀

Estimation Based on Gibbs Sampling

So far, we have argued (without proof) that if we run the Gibbs sampling algorithm for many iterations (through burn-in), we should start to see pairs $(X_1^{(i)}, X_2^{(i)})$ whose joint p.d.f. is nearly the function f from which we wanted to sample. Unfortunately, the successive pairs are not independent of each other even if they do have the same distribution. The law of large numbers does not tell us that the average of dependent random variables with the same distribution converges. However, the type of dependence that we get from a Markov chain is sufficiently regular that there are theorems that guarantee convergence of averages and even that the averages are asymptotically normal. That is, suppose that we wish to estimate the mean μ of some function $h(X_1, X_2)$ based on m observations from the Markov chain. We can still assume that $\frac{1}{m}\sum_{i=1}^{m} h(X_1^{(i)}, X_2^{(i)})$ converges to μ, and that it has approximately the normal distribution with mean μ and variance σ^2/m. However, the convergence will typically be slower than for i.i.d. samples, and σ^2 will be larger than the variance of $h(X_1, X_2)$. The reason for this is that the successive values of $h(X_1^{(i)}, X_2^{(i)})$ are usually positively correlated. The variance of an average of positively correlated identically distributed random variables is higher than the variance of an average of the same number of i.i.d. random variables. (See Exercise 4.)

We shall deal with the problems caused by correlated samples by making use of the same k independent Markov chains that we used for determining how much burn-in to do. Discard the burn-in and continue to sample each Markov chain for m_0 more iterations. From each Markov chain, we compute our desired estimator, either an average, a sample quantile, a sample variance, or other measure, Z_j for $j = 1, \dots, k$. We then compute S as in Eq. (12.2.2); that is,

$$S = \left(\frac{1}{k}\sum_{j=1}^{k}(Z_j - \overline{Z})^2\right)^{1/2}. \tag{12.5.1}$$

Then S^2 is an estimator of the simulation variance of the Z_j's. Write the simulation variance as σ^2/m_0 and estimate σ^2 by $\hat{\sigma}^2 = m_0 S^2$ as we did in Example 12.2.9. Also, combine all samples from all k chains into a single sample, and use this single sample to form the overall estimator Z. The simulation standard error of our estimator Z is then $(\hat{\sigma}^2/(m_0 k))^{1/2} = S/k^{1/2}$.

In addition, we may wish to determine how many simulations to perform in order to obtain a precise estimator. We can substitute $\hat{\sigma}$ for σ in Eq. (12.2.5) to get a proposed number of simulations v. These v simulations would be divided between the k Markov chains so that each chain would be run for at least v/k iterations if $v/k > m_0$.

Some Examples

Example 12.5.3 Nursing Homes in New Mexico. We actually do not need Gibbs sampling in order to simulate a sample from the posterior distribution in Example 12.5.1. The reason is that we have a closed-form expression for the joint distribution of μ and τ in that example. Each of the marginal and conditional distributions are known and easy

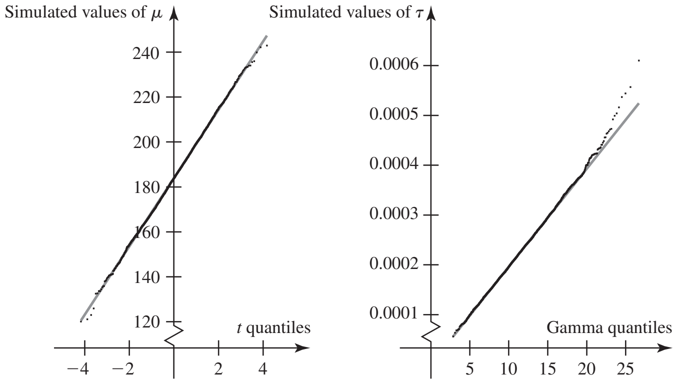


Figure 12.7 Quantile plots of μ and τ values simulated from the posterior distribution in Example 12.5.3. The line on each plot shows the quantiles of the actual posterior distribution as found in Sec. 8.6. The horizontal axis on the left plot is labeled by quantiles of the t distribution with 22 degrees of freedom. The actual posterior of μ is a rescaling and shifting of this t distribution. The horizontal axis on the right plot is labeled by quantiles of the gamma distribution with parameters 11 and 1. The actual posterior of τ is a rescaling of this gamma distribution.

to simulate. Gibbs sampling is most useful when only the conditionals are easy to simulate. However, we can illustrate the use of Gibbs sampling in Example 12.5.1 and compare the simulated results to the known marginal distributions of μ and τ.

In Example 12.5.2, we started $k = 5$ Markov chains and burned them in for 100 iterations. Now we wish to produce a sample of (μ, τ) pairs from the joint posterior distribution. After burn-in, we run another $m_0 = 1000$ iterations for each chain. These iterations produce five correlated sequence of (μ, τ) pairs. The correlations between successive pairs of μ values are quite small. The same is true of successive τ values. To compare the results with the known posterior distributions found in Sec. 8.6, Fig. 12.7 has a t quantile plot of the μ values and a gamma quantile plot of the τ values. (Normal quantile plots were introduced on page 720. Gamma and t quantile plots are constructed in the same way using gamma and t quantile functions in place of the standard normal quantile function.) The simulated values seem to lie close to the lines drawn on the plots in Fig. 12.7. (A few points in the tails stray a bit from the lines, but this occurs with virtually all quantile plots.) The lines in Fig. 12.7 show the quantiles of the actual posterior distributions, which are a t distribution with 22 degrees of freedom multiplied by 15.21 and centered at 183.95 for μ and the gamma distribution with parameters 11 and 50925.37 for τ.

We can use the sample of (μ, τ) pairs to estimate the posterior mean of an arbitrary function of (μ, τ). For example, suppose that we are interested in the mean θ of $\mu + 1.645/\tau^{1/2}$, which is the 0.95 quantile of the unknown distribution of the original observations. The average of our 5000 simulated values of $\mu + 1.645/\tau^{1/2}$ is $Z = 299.67$. The value of S from Eq. (12.5.1) is 0.4119, giving us a value of $\hat{\sigma} = 13.03$. The simulation standard error of Z is then $\hat{\sigma}/5000^{1/2} = 0.1842$. The true posterior mean of $\mu + 1.645/\tau^{1/2}$ can be computed exactly in this example, and it is

$$\mu_1 + 1.645\beta_1^{1/2}\frac{\Gamma(\alpha_1 - .5)}{\Gamma(\alpha_1)} = 299.88,$$

a bit more than 1 simulation standard error away from our simulated value of Z. Suppose that we want our estimator of θ to be within 0.01 of the true value with

probability 0.99. Substituting these values and $\hat{\sigma} = 13.03$ into Eq. (12.2.5), we find that we need $v = 12{,}358{,}425$ total simulations. Each of our five Markov chains would have to be run for 2,251,685 iterations. ◀

The true value of Gibbs sampling begins to emerge in problems with more than two parameters. The general Gibbs sampling algorithm for p random variables $(X_1, \ldots, X_p)$ with p.d.f. $f(\boldsymbol{x}) = cg(\boldsymbol{x})$ is as follows. First, verify that g looks like an easy-to-simulate p.d.f. as a function of each variable for fixed values of all the others. Then perform these steps:

1. Pick starting values $x_2^{(0)}, \ldots, x_p^{(0)}$ for $X_2, \ldots, X_p$, and set $i = 0$.
2. Simulate a new value $x_1^{(i+1)}$ from the conditional distribution of X_1 given $X_2 = x_2^{(i)}, \ldots, X_p = x_p^{(i)}$.
3. Simulate a new value $x_2^{(i+1)}$ from the conditional distribution of X_2 given $X_1 = x_1^{(i+1)}, X_3 = x_3^{(i)}, \ldots, X_p = x_p^{(i)}$.

$\vdots$

$p+1$. Simulate a new value $x_p^{(i+1)}$ from the conditional distribution of X_p given $X_1 = x_1^{(i+1)}, \ldots, X_{p-1} = x_{p-1}^{(i+1)}$.

$p+2$. Replace i by $i+1$, and return to step 2.

The sequence of successive p-tuples of $(X_1, \ldots, X_p)$ values produced by this algorithm is a Markov chain in the same sense as before. The stationary distribution of this Markov chain has the p.d.f. f, and the distribution of an iteration many steps after the start should be approximately the stationary distribution.

Example 12.5.4

Multiple Regression with an Improper Prior. Consider a problem in which we observe data consisting of triples (Y_i, x_{1i}, x_{2i}) for $i = 1, \ldots, n$. We assume that the x_{ji} values are known, and we model the distribution of Y_i as the normal distribution with mean $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ and precision τ. This is the multiple regression model introduced in Sec. 11.5 with the variance replaced by 1 over the precision. Suppose that we use the improper prior $\xi(\beta_0, \beta_1, \beta_2, \tau) = 1/\tau$ for the parameters. The posterior p.d.f. of the parameters is then proportional to the likelihood times $1/\tau$, which is a constant times

$$\tau^{n/2-1} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2\right). \tag{12.5.2}$$

To simplify the ensuing formulas, we shall define some summaries of the data:

$$\overline{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i}, \qquad \overline{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{2i}, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$
$$s_{11} = \sum_{i=1}^{n} x_{1i}^2, \qquad s_{22} = \sum_{i=1}^{n} x_{2i}^2, \qquad s_{12} = \sum_{i=1}^{n} x_{1i}x_{2i},$$
$$s_{1y} = \sum_{i=1}^{n} x_{1i}y_i, \qquad s_{2y} = \sum_{i=1}^{n} x_{2i}y_i, \qquad s_{yy} = \sum_{i=1}^{n} y_i^2.$$

Looking at (12.5.2) as a function of τ for fixed values of β_0, β_1, and β_2, it looks like the p.d.f. of the gamma distribution with parameters $n/2$ and $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2/2$. Looking at (12.5.2) as a function of β_j for fixed values of the other parameters, it is e to the power of a quadratic in β_j with negative coefficient on the

β_j^2 term. As such, it looks like the p.d.f. of a normal random variable with a mean that depends on the data and the other β's and a variance that equals $1/\tau$ times some function of the data. We can be more specific if we complete the square in the expression $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2$ three times, each time treating a different β_j as the variable of interest. For example, treating β_0 as the variable of interest, we get

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2 = n\left(\beta_0 - [\overline{y} - \beta_1 \overline{x}_1 - \beta_2 \overline{x}_2]\right)^2,$$

plus a term that does not depend on β_0. So, the conditional distribution of β_0 given the remaining parameters is the normal distribution with mean $\overline{y} - \beta_1\overline{x}_1 - \beta_2\overline{x}_2$ and variance $1/[n\tau]$. Treating β_1 as the variable of interest, we get

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2 = s_{11}(\beta_1 - w_1)^2,$$

plus a term that does not depend on β_1, where

$$w_1 = \frac{1}{s_{11}}\left(s_{1y} - \beta_0 n\overline{x}_1 - \beta_2 s_{12}\right).$$

This means that the conditional distribution of β_1 given the other parameters is the normal distribution with mean w_1 and variance $(\tau s_{11})^{-1}$. Similarly, the conditional distribution of β_2 given the other parameters is the normal distribution with mean w_2 and variance $(\tau s_{22})^{-1}$, where

$$w_2 = \frac{1}{s_{22}}\left(s_{2y} - \beta_0 n\overline{x}_2 - \beta_1 s_{12}\right).$$ ◀

Example 12.5.5

Unemployment in the 1950s. In Example 11.5.9, we saw that unemployment data from the years 1951–1959 appeared to satisfy the assumptions of the multiple regression model better than the data that included the year 1950. Let us use just the last nine years of data from this example (in Table 11.12). We shall use an improper prior and Gibbs sampling to obtain samples from the posterior distribution of the parameters. The necessary conditional distributions were all given in Example 12.5.4. We just need the values of the summary statistics and $n = 9$:

$$\overline{x}_1 = 140.7778, \quad \overline{x}_2 = 6, \quad \overline{y} = 2.789,$$
$$s_{11} = 179585, \quad s_{22} = 384, \quad s_{12} = 7837,$$
$$s_{1y} = 3580.9, \quad s_{2y} = 169.2, \quad s_{yy} = 78.29.$$

Once again, we shall run $k = 5$ Markov chains. In this problem, there are four coordinates to the parameter: β_i for $i = 0, 1, 2$ and τ. So, we compute four F statistics and burn-in until the largest F is less than $1 + 0.44m$. Suppose that this occurs at $m = 4546$. We then sample 10,000 more iterations from each Markov chain.

Suppose that we want an interval $[a, b]$ that contains 90 percent of the posterior distribution of β_1. The numbers a and b will be the sample 0.05 and 0.95 quantiles. Based on our combined sample of 50,000 values of β_1, the interval is $[-0.1178, -0.0553]$. In order to assess the uncertainty in the endpoints, we compute the 0.05 and 0.95 sample quantiles for each of the five Markov chains. Those values are

0.05 quantiles: $-0.1452, -0.1067, -0.1181, -0.1079, -0.1142$

0.95 quantiles: $-0.0684, -0.0610, -0.0486, -0.0594, -0.0430.$

The value of S based on the sample 0.05 quantiles is 0.01567, and the value of S based on the sample 0.95 quantiles is 0.01142. To be safe, we shall use the larger of these two to estimate the simulation standard errors of our interval endpoints. Since each chain was run for $m_0 =$10,000 iterations, we have $\hat{\sigma} = Sm_0^{1/2} = 1.567$. Suppose that we want each endpoint of the interval to be within 0.01 of the corresponding true quantile of the distribution of β_1 with probability 0.95. (The probability that both endpoints are within 0.01 would be a bit smaller, but is harder to compute.) We could use Eq. (12.2.5) to compute how many simulations we would need. That equation yields $v = 94{,}386$, which means that each of our five chains would need to be run 18,878 iterations, about twice what we already have. For comparison, a 90 percent confidence interval for β_1 constructed using the methods of Sec. 11.3 is $[-0.1124, -0.0579]$. This is quite close to the posterior probability interval. ◀

Although we did not do so in this text, we could have found the posterior distribution for Example 12.5.5 in closed form. Indeed, the 90 percent confidence interval calculated at the end of the example contains 90 percent of the posterior distribution in much the same way that coefficient $1 - \alpha_0$ confidence intervals contain posterior probability $1 - \alpha_0$ in Sec. 11.4 when we use improper priors. The next example is one in which a closed-form solution is not available.

Example 12.5.6 Bayesian Analysis of One-Way Layout with Unequal Variances. Consider the one-way layout that was introduced in Sec. 11.6. There, we assumed that data would be observed from each of p normal distributions with possibly different means but the same variance. In order to illustrate the added power of Gibbs sampling, we shall drop the assumption that each normal distribution has the same variance. That is, for $i = 1, \dots, p$, we shall assume that $Y_{i1}, \dots, Y_{in_i}$ have the normal distribution with mean μ_i and precision τ_i, and all observations are independent conditional on all parameters. Our prior distribution for the parameters will be the following: Let $\mu_1, \dots, \mu_p$ be conditionally independent given all other parameters with μ_i having the normal distribution with mean ψ and precision $\lambda_0\tau_i$. Here, ψ is another parameter that also needs a distribution. We introduce this parameter ψ as a way of saying that we think that the μ_i's all come from a common distribution, but we are not willing to say for sure where that distribution is located. We then say that ψ has the normal distribution with mean ψ_0 and precision u_0. For an improper prior, we could set $u_0 = 0$ in what follows, and then ψ_0 would not be needed either. Next, we model $\tau_1, \dots, \tau_p$ as i.i.d. having the gamma distribution with parameters α_0 and β_0. We model ψ and the τ_i's as independent. For an improper prior, we could set $\alpha_0 = \beta_0 = 0$. The type of model just described is called a *hierarchical model* because of the way that the distributions fall into a hierarchy of levels. Figure 12.8 illustrates the levels of the hierarchy in this example.

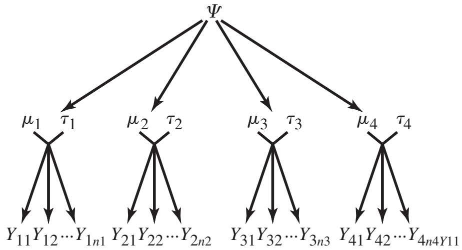


Figure 12.8 Diagram of hierarchical model in Example 12.5.6. The parameter ψ influences the distributions of the μ_i's, while the (μ_i, τ_i) parameters influence the distributions of the Y_{ij}'s.

The joint p.d.f. of the observations and the parameters is the product of the likelihood function (the p.d.f. of the observations given the μ_i's and τ_i's) times the product of the conditional prior p.d.f.'s of the μ_i's given the τ_i's and ψ, times the prior p.d.f.'s of the τ_i's times the prior p.d.f. for ψ. Aside from constants that depend neither on the data nor on the parameters, this product has the form

$$\exp\left(-\frac{u_0(\psi-\psi_0)^2}{2}-\sum_{i=1}^{p}\tau_i\left[\beta_0+\frac{n_i(\mu_i-\overline{y}_i)^2+w_i+\lambda_0(\mu_i-\psi)^2}{2}\right]\right)$$
$$\times\prod_{i=1}^{p}\tau_i^{\alpha_0+[n_i+1]/2-1}, \tag{12.5.3}$$

where $w_i=\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2$ for $i=1,\ldots,p$. We have arranged terms in (12.5.3) so that the terms involving each parameter are close together. This will facilitate describing the Gibbs sampling algorithm.

In order to set up Gibbs sampling, we need to examine (12.5.3) as a function of each parameter separately. The parameters are $\mu_1,\ldots,\mu_p$; $\tau_1,\ldots,\tau_p$; and ψ. As a function of τ_i, (12.5.3) looks like the p.d.f. of the gamma distribution with parameters $\alpha_0+(n_i+1)/2$ and $\beta_0+[n_i(\mu_i-\overline{y}_i)^2+w_i+\lambda_0(\mu_i-\psi)^2]/2$. As a function of ψ, it looks like the p.d.f. of the normal distribution with mean $[u_0\psi_0+\lambda_0\sum_{i=1}^{p}\tau_i\mu_i]/[u_0+\lambda_0\sum_{i=1}^{p}\tau_i]$ and precision $u_0+\lambda_0\sum_{i=1}^{p}\tau_i$. This is obtained by completing the square for all terms involving ψ. Similarly, by completing the square for all terms involving μ_i, we find that (12.5.3) looks like the normal p.d.f. with mean $[n_i\overline{y}_i+\lambda_0\psi]/[n_i+\lambda_0]$ and precision $\tau_i(n_i+\lambda_0)$ as a function of μ_i. All of these distributions are easy to simulate.

As an example, use the hot dog calorie data from Example 11.6.2. In this example, $p=4$. We shall use a prior distribution in which $\lambda_0=\alpha_0=1$, $\beta_0=0.1$, $u_0=0.001$, and $\psi_0=170$. We use $k=6$ Markov chains and do $m=100$ burn-in simulations, which turn out to be more than enough to make the maximum of all nine F statistics less than $1+0.44m$. We then run each of the six Markov chains another 10,000 iterations. The samples from the posterior distribution allow us to answer any questions that we might have about the parameters, including some that we would not have been able to answer using the analysis done in Chapter 11. For example, the posterior means and standard deviations of some of the parameters are listed in Table 12.6. Notice how much different the variances $1/\tau_i$ seem to be in the four groups. For example, we can compute the probability that $1/\tau_4>4/\tau_1$ by counting up the number of iterations ℓ in which $1/\tau_4^{(\ell)}>4/\tau_1^{(\ell)}$ and dividing by 60,000. The result is 0.5087, indicating that there is a large chance that at least some of the variances are quite different. If the variances are different, the ANOVA calculations in Chapter 11 are not justified.

We can also address the question of how much difference there is between the μ_i's. For comparison, we shall do the same calculations that we did in Example 12.3.7. In 99 percent of the 60,000 simulations, at least one $|\mu_i^{(\ell)}-\mu_j^{(\ell)}|>24.66$. In about one-half of the simulations, all $|\mu_i^{(\ell)}-\mu_j^{(\ell)}|>2.268$. And in 99 percent of the simulations, the average of the differences was at least 13.07. Figure 12.9 contains a plot of the sample c.d.f.'s of the largest, smallest, and average of the six $|\mu_i-\mu_j|$ differences. Careful examination of the results in this example shows that the four μ_i's appear to be closer together than we would have thought after the analysis of Example 12.3.7. This is typical of what occurs when we use a proper prior in a hierarchical model. In Example 12.3.7, the μ_i's were all independent, and they did not have a common unknown mean in the prior. In Example 12.5.6, the μ_i's all have a common prior distribution with mean ψ, which is an additional unknown parameter. The estimation

Table 12.6 Posterior means and standard deviations for some parameters in Example 12.5.6

Type	Beef	Meat	Poultry	Specialty
i	1	2	3	4
$E(\mu_i\|\boldsymbol{y})$	156.7	158.4	120.7	159.7
$(Var(\mu_i\|\boldsymbol{y}))^{1/2}$	3.498	5.241	6.160	10.55
$E(1/\tau_i\|\boldsymbol{y})$	252.3	487.3	670.8	1100
$(Var(1/\tau_i\|\boldsymbol{y}))^{1/2}$	84.70	179.1	250.6	586.9
$E(\psi\|\boldsymbol{y}) = 152.8$		$(Var(\psi\|\boldsymbol{y}))^{1/2} = 10.42$		

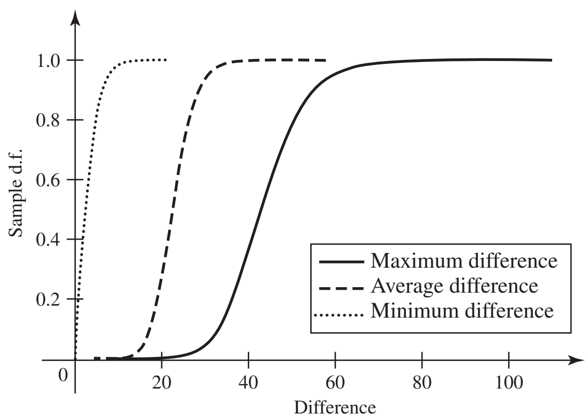


Figure 12.9 Sample c.d.f.'s of the maximum, average, and minimum of the six $|\mu_i - \mu_j|$ differences for Example 12.5.6.

of this additional parameter allows the posterior distributions of the μ_i's to be pulled toward a location that is near the average of all of the samples. With these data, the overall sample average is 147.60. ◀

Prediction

All of the calculations done in the examples of this section have concerned functions of the parameters. The sample from the posterior distribution that we obtain from Gibbs sampling can also be used to make predictions and form prediction intervals for future observations. The most straightforward way to make predictions is to simulate the future data conditional on each value of the parameter from the posterior sample. Although there are more efficient methods for predicting, this method is easy to describe and evaluate.

Example 12.5.7 Calories in Hot Dogs. In Example 12.5.6, we might be concerned with how different we should expect the calorie counts of two hot dogs to be. For example, let Y_1 and Y_3 be future calorie counts for hot dogs of the beef and poultry varieties, respectively. We can form a prediction interval for $D = Y_1 - Y_3$ as follows: For each iteration ℓ, let

the simulated parameter vector be

$$\theta^{(\ell)} = \left(\mu_1^{(\ell)}, \mu_2^{(\ell)}, \mu_3^{(\ell)}, \mu_4^{(\ell)}, \tau_1^{(\ell)}, \tau_2^{(\ell)}, \tau_3^{(\ell)}, \tau_4^{(\ell)}, \psi^{(\ell)}, \beta^{(\ell)}\right).$$

For each ℓ, simulate a beef hot dog calorie count $Y_1^{(\ell)}$ having the normal distribution with mean $\mu_1^{(\ell)}$ and variance $1/\tau_1^{(\ell)}$. Also simulate a poultry hot dog calorie count $Y_3^{(\ell)}$ having the normal distribution with mean $\mu_3^{(\ell)}$ and variance $1/\tau_3^{(\ell)}$. Then compute $D^{(\ell)} = Y_1^{(\ell)} - Y_3^{(\ell)}$. Sample quantiles of the values $D^{(1)}, \dots, D^{(60,000)}$ can be used to estimate quantiles of the distribution of D.

For example, suppose that we want a 90 percent prediction interval for D. We simulate 60,000 $D^{(\ell)}$ values as above and find the 0.05 and 0.95 sample quantiles to be −14.86 and 87.35, which are then the endpoints of our prediction interval. To assess how close the simulation estimators are to the actual quantiles of the distribution of D, we compute the simulation standard errors of the two endpoints. For the samples from each of the $k = 6$ Markov chains, we can compute the sample 0.05 quantiles of our D values. We can then use these values as $Z_1, \dots, Z_6$ in Eq. (12.5.1) to compute a value S. Our simulation standard error is then $S/6^{1/2}$. We can then repeat this for the sample 0.95 quantiles. For the two endpoints of our interval, the simulation standard errors are 0.2447 and 0.3255, respectively. These simulation standard errors are fairly small compared to the length of the prediction interval. ◄

Example 12.5.8

Censored Arsenic Measurements. Frey and Edwards (1997) describe the National Arsenic Occurrence Survey (NAOS). Several hundred community water systems submitted samples of their untreated water in an attempt to help characterize the distribution of arsenic across the nation. Arsenic is one of several contaminants that the Environmental Protection Agency (EPA) is required to regulate. One difficulty in modeling the occurrence of a substance like arsenic is that concentrations are often too low to be measured accurately. In such cases, the measurements are censored. That is, we only know that the concentration of arsenic is less than some censoring point, but not how much less. In the NAOS data set, the censoring point is 0.5 microgram per liter. Each concentration less than 0.5 microgram per liter is censored.

Gibbs sampling can help us to estimate the distribution of arsenic in spite of the censored observations. Lockwood et al. (2001) do an extensive analysis of the NAOS and other data and show how the distribution of arsenic differs from one state to the next and from one type of water source to the next. For convenience, let us focus our attention on the 24 observations from one state, Ohio. Of those 24 observations, 11 were taken from groundwater sources (wells). The other 13 came from surface water sources (e.g., rivers and lakes). The following are seven uncensored groundwater observations from Ohio:

9.62, 10.50, 2.30, 0.80, 17.04, 9.90, 1.32.

The other four groundwater observations were censored.

Suppose that we model groundwater arsenic concentrations in Ohio as having the lognormal distribution with parameters μ and σ^2. One popular way to deal with censored observations is to treat them like unknown parameters. That is, let $Y_1, \dots, Y_4$ be the four unknown concentrations from the four wells where the measurements were censored. Let $X_1, \dots, X_7$ stand for the seven uncensored values. Suppose that μ and $\tau = 1/\sigma^2$ have the normal-gamma prior distribution with hyperparameters μ_0, λ_0, α_0, and β_0. The joint p.d.f. of $X_1, \dots, X_7, Y_1, \dots, Y_4$, and μ and τ is proportional to

$$\tau^{\beta_0+(7+4+1)/2-1}\exp\left(-\frac{\tau}{2}\left[\lambda_0(\mu-\mu_0)^2+\sum_{i=1}^{7}(\log(x_i)-\mu)^2+\sum_{j=1}^{4}(\log(y_i)-\mu)^2+2\beta_0\right]\right).$$

The observed data consist of the values $x_1, \ldots, x_7$ of $X_1, \ldots, X_7$ together with the fact that $Y_j \le 0.5$ for $j = 1, \ldots, 4$. The conditional distributions of μ and τ given the data and the other parameters are just like what we obtained in Example 12.5.1. To be precise, μ has the normal distribution with mean

$$\frac{\lambda_0\mu_0+\sum_{i=1}^{7}\log(x_i)+\sum_{j=1}^{4}\log(y_j)}{\lambda_0+11}$$

and precision $\tau(\lambda_0 + 11)$ conditional on τ, the Y_j's, and the data. Also, τ has the gamma distribution with parameters $\alpha_0 + (11 + 1)/2$ and

$$\beta_0+\frac{1}{2}\left(\sum_{i=1}^{7}(\log(x_i)-\mu)^2+\sum_{j=1}^{4}(\log(y_i)-\mu)^2+\lambda_0(\mu-\mu_0)^2\right),$$

conditional on μ, the Y_j's, and the data. The conditional distribution of the Y_j's given μ, τ, and the data is that of i.i.d. random variables with the lognormal distribution having parameters μ and $1/\tau$ but conditional on $Y_j < 0.5$. That is, the conditional c.d.f. of each Y_j is

$$F(y)=\frac{\Phi([\log(y)-\mu]\tau^{1/2})}{\Phi([\log(0.5)-\mu]\tau^{1/2})},\quad \text{for } y<0.5.$$

We can simulate random variables with c.d.f. F so long as we can compute the standard normal c.d.f. and quantile function. Let U have the uniform distribution on the interval [0, 1]. Then

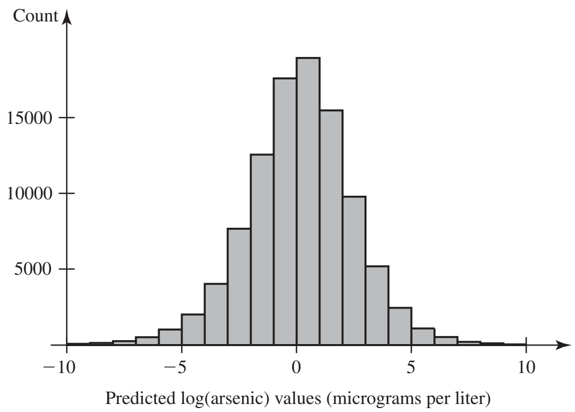
$$Y=\exp\big(\mu+\tau^{-1/2}\Phi^{-1}[U\Phi([\log(0.5)-\mu]\tau^{1/2})]\big)$$

has the desired c.d.f., F.

One example of the type of inference that is needed in an analysis of this sort is to predict arsenic concentrations for different water systems. Knowing the likely sizes of arsenic measurements can help water systems choose economical treatments that will meet the standards set by the EPA. For simplicity, we shall simulate one arsenic concentration at each iteration of the Markov chain. For example, suppose that $(\mu^{(i)}, \tau^{(i)})$ are the simulated values of μ and τ at the ith iteration of the Markov chain. Then we can simulate $Y^{(i)} = \exp(\mu^{(i)} + Z(\tau^{(i)})^{-1/2})$, where Z is a standard normal random variable. Figure 12.10 shows a histogram of the simulated $\log(Y^{(i)})$ values from 10 Markov chains of length 10,000 each. The proportion of predicted values that are below the censoring point of log(0.5) is 0.335, with a simulation standard error of 0.001. The median predicted value on the logarithmic scale is 0.208 with a simulation standard error of 0.007. We can transform this back to the original scale of measurement using the delta method as described in Example 12.2.8. The median predicted arsenic concentration is $\exp(0.208) = 1.231$ micrograms per liter with a simulation standard error of $0.007\exp(0.208) = 0.009$. ◄

Note: There Are More-General Markov Chain Monte Carlo Algorithms. Gibbs sampling requires a special structure for the distribution we wish to simulate. We need to be able to simulate the conditional distribution of each coordinate given the other coordinates. In many problems, this is not possible for at least some, if not all, of the coordinates. If only one coordinate is difficult to simulate, one might try using an acceptance/rejection simulator for that one coordinate. If even this does not work,

Figure 12.10 Histogram of simulated log(arsenic) values for 10,000 iterations from each of 10 Markov chains in Example 12.5.8. The vertical line is at the censoring point, log(0.5).



there are more-general Markov chain Monte Carlo algorithms that can be used. The simplest of these is the *Metropolis algorithm* introduced by Metropolis et al. (1953). An introduction to the Metropolis algorithm can be found in chapter 11 of Gelman et al. (1995) together with a further generalization due to Hastings (1970).

Summary

We introduced the Gibbs sampling algorithm that produces a Markov chain of observations from a joint distribution of interest. The joint distribution must have a special form. As a function of each variable, the joint p.d.f. must look like a p.d.f. from which it is easy to simulate pseudo-random variables. The Gibbs sampling algorithm cycles through the coordinates, simulating each one conditional on the values of the others. The algorithm requires a burn-in period during which the distribution of states in the Markov chain converges to the desired distribution. Assessing convergence and computing simulation standard errors of simulated values are both facilitated by running several independent Markov chains simultaneously.

Exercises

1. Let $f(x_1, x_2) = cg(x_1, x_2)$ be a joint p.d.f. for (X_1, X_2). For each x_2, let $h_2(x_1) = g(x_1, x_2)$. That is, h_2 is what we get by considering $g(x_1, x_2)$ as a function of x_1 for fixed x_2. Show that there is a multiplicative factor c_2 that does not depend on x_1 such that $h_2(x_1)c_2$ is the conditional p.d.f. of X_1 given $X_2 = x_2$.

2. Let $f(x_1, x_2)$ be a joint p.d.f. Suppose that $(X_1^{(i)}, X_2^{(i)})$ has the joint p.d.f. f. Let $(X_1^{(i+1)}, X_2^{(i+1)})$ be the result of applying steps 2 and 3 of the Gibbs sampling algorithm on page 824. Prove that $(X_1^{(i+1)}, X_2^{(i)})$ and $(X_1^{(i+1)}, X_2^{(i+1)})$ also have the joint p.d.f. f.

3. Let $Z_1, Z_2, \ldots$ form a Markov chain, and assume that the distribution of Z_1 is the stationary distribution. Show that the joint distribution of (Z_1, Z_2) is the same as the joint distribution of (Z_i, Z_{i+1}) for all $i > 1$. For convenience, you may assume that the Markov chain has finite state space, but the result holds in general.

4. Let $X_1, \ldots, X_n$ be uncorrelated, each with variance σ^2. Let $Y_1, \ldots, Y_n$ be positively correlated, each with variance σ^2. Prove that the variance of $\overline{X}$ is smaller than the variance of $\overline{Y}$.

5. Use the data consisting of 30 lactic acid concentrations in cheese, 10 from Example 8.5.4 and 20 from Exercise 16 in Sec. 8.6. Fit the same model used in Example 8.6.2 with the same prior distributon, but this time use the Gibbs sampling algorithm described in Example 12.5.1. Simulate 10,000 pairs of (μ, τ) parameters. Estimate the posterior mean of $(\sqrt{\tau}\mu)^{-1}$, and compute the simulation standard error of the estimator.

6. Use the data on dishwasher shipments in Table 11.13 on page 744. Suppose that we wish to fit a multiple linear regression model for predicting dishwasher shipments from time (year minus 1960) and private residential investment. Suppose that the parameters have the improper prior proportional to $1/\tau$. Use the Gibbs sampling algorithm to obtain a sample of size 10,000 from the joint posterior distribution of the parameters.

a. Let β_1 be the coefficient of time. Draw a plot of the sample c.d.f. of $|\beta_1|$ using your posterior sample.

b. We are interested in predicting dishwasher shipments for 1986.

i. Draw a histogram of the values of $\beta_0 + 26\beta_1 + 67.2\beta_2$ from your posterior distribution.

ii. For each of your simulated parameters, simulate a dishwasher sales figure for 1986 (time = 26 and private residential investment = 67.2). Compute a 90 percent prediction interval from the simulated values and compare it to the interval found in Example 11.5.7.

iii. Draw a histogram of the simulated 1986 sales figures, and compare it to the histogram in part i. Can you explain why one sample seems to have larger variance than the other?

7. Use the data in Table 11.19 on page 762. This time fit the model developed in Example 12.5.6. Use the prior hyperparameters $\lambda_0 = \alpha_0 = 1, \beta_0 = 0.1, u_0 = 0.001$, and $\psi_0 = 800$. Obtain a sample of 10,000 from the posterior joint distribution of the parameters. Estimate the posterior means of the three parameters μ_1, μ_2, and μ_3.

8. In this problem, we shall outline a form of robust linear regression. Assume throughout the exercise that the data consist of pairs (Y_i, x_i) for $i = 1, \dots, n$. Assume also that the x_i's are all known and the Y_i's are independent random variables. We shall only deal with simple regression here, but the method easily extends to multiple regression.

a. Let β_0, β_1, and σ stand for unknown parameters, and let a be a known positive constant. Prove that the following two models are equivalent. That is, prove that the joint distribution of $(Y_1, \dots, Y_n)$ is the same in both models.

Model 1: For each i, $[Y_i - (\beta_0 + \beta_1 x_i)]/\sigma$ has the t distribution with a degrees of freedom.

Model 2: For each i, Y_i has the normal distribution with mean $\beta_0 + \beta_1 x_i$ and variance $1/\tau_i$ conditional on τ_i. Also, $\tau_1, \dots, \tau_n$ are i.i.d. having the gamma distribution with parameters $a/2$ and $a\sigma^2/2$.

Hint: Use the same argument that produced the marginal distribution of μ in Sec. 8.6 when μ and τ had a normal-gamma distribution.

b. Now consider Model 2 from part (a). Let $\eta = \sigma^2$, and assume that η has a prior distribution that is the gamma distribution with parameters $b/2$ and $f/2$, where b and f are known constants. Assume that the parameters β_0 and β_1 have an improper prior with "p.d.f." 1. Show that the product of likelihood and prior "p.d.f." is a constant times

$$\eta^{(na+b)/2-1} \prod_{i=1}^{n} \tau_i^{(a+1)/2-1} \exp\Bigg(-\frac{1}{2}\bigg[f\eta + \sum_{i=1}^{n} \tau_i \left\{a\eta + (y_i - \beta_0 - \beta_1 x_i)^2\right\}\bigg]\Bigg). \quad (12.5.4)$$

c. Consider (12.5.4) as a function of each parameter for fixed values of the others. Show that Table 12.7 specifies the appropriate conditional distribution for each parameter given all of the others.

Table 12.7 Parameters and conditional distributions for Exercise 8

Parameter	(12.5.4) looks like the p.d.f. of this distribution
η	gamma distribution with parameters $(na+b)/2$ and $\left(f + a\sum_{i=1}^n \tau_i\right)/2$
τ_i	gamma distribution with parameters $(a+1)/2$ and $[a\eta + (y_i - \beta_0 - \beta_1 x_i)^2]/2$
β_0	normal distribution with mean $\sum_{i=1}^n \tau_i(y_i - \beta_1 x_i)/\sum_{i=1}^n \tau_i$ and precision $\sum_{i=1}^n \tau_i$
β_1	normal distribution with mean $\sum_{i=1}^n \tau_i x_i(y_i - \beta_0)/\sum_{i=1}^n \tau_i x_i^2$ and precision $\sum_{i=1}^n \tau_i x_i^2$

9. Use the data in Table 11.5 on page 699. Suppose that Y_i is the logarithm of pressure and x_i is the boiling point for the ith observation, $i = 1, \dots, 17$. Use the robust regression scheme described in Exercise 8 with $a = 5$, $b = 0.1$, and $f = 0.1$. Estimate the posterior means and standard deviations of the parameters β_0, β_1, and η.

10. In this problem, we shall outline a Bayesian solution to the problem described in Example 7.5.10 on page 423. Let $\tau = 1/\sigma^2$ and use a proper normal-gamma prior of the form described in Sec. 8.6. In addition to the two parameters μ and τ, introduce n additional parameters.

For $i = 1, \dots, n$, let $Y_i = 1$ if X_i came from the normal distribution with mean μ and precision τ, and let $Y_i = 0$ if X_i came from the standard normal distribution.

a. Find the conditional distribution of μ given τ; $Y_1, \dots, Y_n$; and $X_1, \dots, X_n$.

b. Find the conditional distribution of τ given μ; $Y_1, \dots, Y_n$; and $X_1, \dots, X_n$.

c. Find the conditional distribution of Y_i given μ; τ; $X_1, \dots, X_n$; and the other Y_j's.

d. Describe how to find the posterior distribution of μ and τ using Gibbs sampling.

e. Prove that the posterior mean of Y_i is the posterior probability that X_i came from the normal distribution with unknown mean and variance.

11. Consider, once again, the model described in Example 7.5.10. Assume that $n = 10$ and the observed values of $X_1, \ldots, X_{10}$ are

$-0.92, -0.33, -0.09, 0.27, 0.50, -0.60, 1.66, -1.86,$
$3.29, 2.30.$

a. Fit the model to the observed data using the Gibbs sampling algorithm developed in Exercise 10. Use the following prior hyperparameters: $\alpha_0 = 1$, $\beta_0 = 1$, $\mu_0 = 0$, and $\lambda_0 = 1$.

b. For each i, estimate the posterior probability that X_i came from the normal distribution with unknown mean and variance.

12. Let $X_1, \ldots, X_n$ be i.i.d. with the normal distribution having mean μ and precision τ. Gibbs sampling allows one to use a prior distribution for (μ, τ) in which μ and τ are independent. Let the prior distribution of μ be the normal distribution with mean μ_0 and variance γ_0. Let the prior distribution of τ be the gamma distribution with parameters α_0 and β_0.

a. Show that Table 12.8 specifies the appropriate conditional distribution for each parameter given the other.

b. Use the New Mexico nursing home data (Examples 12.5.2 and 12.5.3). Let the prior hyperparameters be $\alpha_0 = 2$, $\beta_0 = 6300$, $\mu_0 = 200$, and $\gamma_0 = 6.35 \times 10^{-4}$. Implement a Gibbs sampler to find the posterior distribution of (μ, τ). In particular, calculate an interval containing 95 percent of the posterior distribution of μ.

Table 12.8 Parameters and conditional distributions for Exercise 12

Parameter	Prior times likelihood looks like the p.d.f. of this distribution
τ	gamma distribution with parameters $\alpha_0 + n/2$ and $\beta_0 + 0.5\sum_{i=1}^n (x_i - \overline{x})^2 + 0.5n(\overline{x} - \mu)^2$,
μ	normal distribution with mean $(\gamma_0\mu_0 + n\tau\overline{x})/(\gamma_0 + n\tau)$ and precision $\gamma_0 + n\tau$.

13. Consider again the situation described in Exercise 12. This time, we shall let the prior distribution of μ be more like it was in the conjugate prior. Introduce another parameter γ, whose prior distribution is the gamma distribution with parameters a_0 and b_0. Let the prior distribution of μ conditional on γ be the normal distribution with mean μ_0 and precision γ.

a. Prove that the marginal prior distribution of μ specifies that

$$\left(\frac{b_0}{a_0}\right)^{1/2}(\mu - \mu_0) \text{ has the } t \text{ distribution with } 2a_0 \text{ degrees of freedom.}$$

Hint: Look at the derivation of the marginal distribution of μ in Sec. 8.6.

b. Suppose that we want the marginal prior distributions of both μ and τ to be the same as they were with the conjugate prior in Sec. 8.6. How must the prior hyperparameters be related in order to make this happen?

c. Show that Table 12.9 specifies the appropriate conditional distribution for each parameter given the others.

Table 12.9 Parameters and conditional distributions for Exercise 13

Parameter	Prior times likelihood looks like the p.d.f. of this distribution
τ	gamma distribution with parameters $\alpha_0 + n/2$ and $\beta_0 + 0.5\sum_{i=1}^n (x_i - \overline{x})^2 + 0.5n(\overline{x} - \mu)^2$,
μ	normal distribution with mean $(\gamma\mu_0 + n\tau\overline{x})/(\gamma + n\tau)$ and precision $\gamma + n\tau$,
γ	gamma distribution with parameters $a_0 + 1/2$ and $b_0 + 0.5(\mu - \mu_0)^2$.

d. Use the New Mexico nursing home data (Examples 12.5.2 and 12.5.3). Let the prior hyperparameters be $\alpha_0 = 2$, $\beta_0 = 6300$, $\mu_0 = 200$, $a_0 = 2$, and $b_0 = 3150$. Implement a Gibbs sampler to find the posterior distribution of (μ, τ, γ). In particular, calculate an interval containing 95 percent of the posterior distribution of μ.

14. Consider the situation described in Example 12.5.8. In addition to the 11 groundwater sources, there are 13 observations taken from surface water sources in Ohio. Of the 13 surface water measurements, only one was censored. The 12 uncensored surface water arsenic concentrations from Ohio are

1.93, 0.99, 2.21, 2.29, 1.15, 1.81, 2.26, 3.10, 1.18, 1.00,
2.67, 2.15.

a. Fit the same model as described in Example 12.5.8, and predict a logarithm of surface water concentration for each iteration of the Markov chain.

b. Compare a histogram of your predicted measurements to the histogram of the underground well predictions in Fig. 12.10. Describe the main differences.

c. Estimate the median of the distribution of predicted surface water arsenic concentration and compare it to the median of the distribution of predicted groundwater concentration.

15. Let $X_1, \dots, X_{n+m}$ be a random sample from the exponential distribution with parameter θ. Suppose that θ has the gamma prior distribution with known parameters α and β. Assume that we get to observe $X_1, \dots, X_n$, but $X_{n+1}, \dots, X_{n+m}$ are censored.

a. First, suppose that the censoring works as follows: For $i = 1, \dots, m$, if $X_{n+i} \le c$, then we learn only that $X_{n+i} \le c$, but not the precise value of X_{n+i}. Set up a Gibbs sampling algorithm that will allow us to simulate the posterior distribution of θ in spite of the censoring.

b. Next, suppose that the censoring works as follows: For $i = 1, \dots, m$, if $X_{n+i} \ge c$, then we learn only that $X_{n+i} \ge c$, but not the precise value of X_{n+i}. Set up a Gibbs sampling algorithm that will allow us to simulate the posterior distribution of θ in spite of the censoring.

16. Suppose that the time to complete a task is the sum of two parts X and Y. Let (X_i, Y_i) for $i = 1, \dots, n$ be a random sample of the times to complete the two parts of the task. However, for some observations, we get to observe only $Z_i = X_i + Y_i$. To be precise, suppose that we observe (X_i, Y_i) for $i = 1, \dots, k$ and we observe Z_i for $i = k+1, \dots, n$. Suppose that all X_i and Y_j are independent with every X_i having the exponential distribution with parameter λ and every Y_j having the exponential distribution with parameter μ.

a. Prove that the conditional distribution of X_i given $Z_i = z$ has the c.d.f.

$$G(x|z) = \frac{1 - \exp(-x[\lambda - \mu])}{1 - \exp(-z[\lambda - \mu])}, \quad \text{for } 0 < x < z.$$

b. Suppose that the prior distribution of (λ, μ) is as follows: The two parameters are independent with λ having the gamma distribution with parameters a and b, and μ having the gamma distribution with parameters c and d. The four numbers a, b, c, and d are all known constants. Set up a Gibbs sampling algorithm that allows us to simulate the posterior distribution of (λ, μ).

12.6 The Bootstrap

The parametric and nonparametric bootstraps are methods for replacing an unknown distribution F with a known distribution in a probability calculation. If we have a sample of data from the distribution F, we first approximate F by $\hat{F}$ and then perform the desired calculation. If $\hat{F}$ is a good approximation to F, the bootstrap can be successful. If the desired calculation is sufficiently difficult, we typically resort to simulation.

Introduction

Assume that we have a sample $\boldsymbol{X} = (X_1, \dots, X_n)$ of data from some unknown distribution F. Suppose that we are interested in some quantity that involves both F and $\boldsymbol{X}$, for example, the bias of a statistic $g(\boldsymbol{X})$ as an estimator of the median of F. The main idea behind bootstrap analysis in the simplest cases is the following: First, replace the unknown distribution F with a known distribution $\hat{F}$. Next, let $\boldsymbol{X}^*$ be a sample from the distribution $\hat{F}$. Finally, compute the quantity of interest based on $\hat{F}$ and $\boldsymbol{X}^*$, for example, the bias of $g(\boldsymbol{X}^*)$ as an estimator of the median of $\hat{F}$. Consider the following overly simple example.

Example 12.6.1

The Variance of the Sample Mean. Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random sample from a distribution with a continuous c.d.f. F. For the moment, we shall assume nothing more about F than that it has finite mean μ and finite variance σ^2. Suppose that we are interested in the variance of the sample mean $\overline{X}$. We already know that this variance equals σ^2/n, but we do not know σ^2. In order to estimate σ^2/n, the bootstrap replaces

the unknown distribution F with a known distribution $\hat{F}$, which also has finite mean $\hat{\mu}$ and finite variance $\hat{\sigma}^2$. If $\boldsymbol{X}^* = (X_1^*, \dots, X_n^*)$ is a random sample from $\hat{F}$, then the variance of the sample mean $\overline{X}^* = \frac{1}{n}\sum_{i=1}^n X_i^*$ is $\hat{\sigma}^2/n$. Since the distribution $\hat{F}$ is known, we should be able to compute $\hat{\sigma}^2/n$, and we can then use this value to estimate σ^2/n.

One popular choice of known distribution $\hat{F}$ is the sample c.d.f. F_n defined in Sec. 10.6. This sample c.d.f. is the discrete c.d.f. that has jumps of size $1/n$ at each of the observed values $x_1, \dots, x_n$ of the random sample $X_1, \dots, X_n$. So, if $\boldsymbol{X}^* = (X_1^*, \dots, X_n^*)$ is a random sample from $\hat{F}$, then each X_i^* is a discrete random variable with the p.f.

$$f(x) = \begin{cases} \frac{1}{n} & \text{if } x \in \{x_1, \dots, x_n\}, \\ 0 & \text{otherwise.} \end{cases}$$

It is relatively simple to compute the variance $\hat{\sigma}^2$ of a random variable X_i^* whose p.f. is f. The variance is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2,$$

where $\overline{x}$ is the average of the observed values $x_1, \dots, x_n$. Thus, our bootstrap estimate of the variance of $\overline{X}$ is $\hat{\sigma}^2/n$. ◄

The key step in a bootstrap analysis is the choice of the known distribution $\hat{F}$. The particular choice made in Example 12.6.1, namely, the sample c.d.f., leads to what is commonly called the *nonparametric bootstrap*. The reason for this name is that we do not assume that the distribution belongs to a parametric family when choosing $\hat{F} = F_n$. If we are willing to assume that F belongs to a parametric family, then we can choose $\hat{F}$ to be a member of that family and perform a *parametric bootstrap* analysis as illustrated next.

Example 12.6.2 **The Variance of the Sample Mean.** Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random sample from the normal distribution with mean μ and variance σ^2. Suppose, as in Example 12.6.1, that we are interested in estimating σ^2/n, the variance of the sample mean $\overline{X}$. To apply the parametric bootstrap, we replace F by $\hat{F}$, a member of the family of normal distributions. For this example, we shall choose $\hat{F}$ to be the normal distribution with mean and variance equal to the M.L.E.'s $\overline{x}$ and $\hat{\sigma}^2$, respectively, although other choices could be made. We then estimate σ^2/n by the variance of the sample mean $\overline{X}^*$ of a random sample from the distribution $\hat{F}$. The variance of $\overline{X}^*$ is easily computed as $\hat{\sigma}^2/n$. In this case, the parametric bootstrap yields precisely the same answer as the nonparametric bootstrap. ◄

In Examples 12.6.1 and 12.6.2, it was very simple to compute the variance of the sample mean of a random sample from the distribution $\hat{F}$. In typical applications of the bootstrap, it is not so simple to compute the quantity of interest. For example, there is no simple formula for the variance of the sample median of a sample $\boldsymbol{X}^*$ from $\hat{F}$ in Examples 12.6.1 and 12.6.2. In such cases, one resorts to simulation techniques in order to approximate the desired calculation. Before presenting examples of the use of simulation in the bootstrap, we shall first describe the general class of situations in which bootstrap analysis is used.

Table 12.10 Correspondence between statistical model and bootstrap analysis

	Statistical model	Bootstrap
Distribution	Unknown F	Known $\hat{F}$
Data	i.i.d. sample $\boldsymbol{X}$ from F	i.i.d. sample $\boldsymbol{X}^*$ from $\hat{F}$
Function of interest	$\eta(\boldsymbol{X}, F)$	$\eta(\boldsymbol{X}^*, \hat{F})$
Parameter/estimate	Mean, median, variance, etc. of $\eta(\boldsymbol{X}, F)$	Mean, median, variance, etc. of $\eta(\boldsymbol{X}^*, \hat{F})$

The Bootstrap in General

Let $\eta(\boldsymbol{X}, F)$ be a quantity of interest that possibly depends on both a distribution F and a sample $\boldsymbol{X}$ drawn from F. For example, if the distribution F has the p.d.f. f, we might be interested in

$$\eta(\boldsymbol{X}, F) = \left[\frac{1}{n}\sum_{i=1}^{n} X_i - \int x f(x)\, dx\right]^2. \tag{12.6.1}$$

In Examples 12.6.1 and 12.6.2, we wanted the variance of the sample average, which equals the mean of the quantity in Eq. (12.6.1). In general, we might wish to estimate the mean or a quantile or some other probabilistic feature of $\eta(\boldsymbol{X}, F)$. The bootstrap estimates the mean or a quantile or some other feature of $\eta(\boldsymbol{X}, F)$ by the mean or quantile or the other feature of $\eta(\boldsymbol{X}^*, \hat{F})$, where $\boldsymbol{X}^*$ is a random sample drawn from the distribution $\hat{F}$, and $\hat{F}$ is some distribution that we hope is close to F. Table 12.10 shows the correspondence between the original statistical model for the data and the quantities that are involved in a bootstrap analysis. The function η of interest must be something that exists for all distributions under consideration and all samples from those distributions. Other quantities that might be of interest include the quantiles of the distribution of a statistic, the M.A.E. or M.S.E. of an estimator, the bias of an estimator, probabilities that statistics lie in various intervals, and the like.

In the simple examples considered so far, the distribution of $\eta(\boldsymbol{X}^*, \hat{F})$ was both known and easy to compute. It will often be the case that the distribution of $\eta(\boldsymbol{X}^*, \hat{F})$ is too complicated to allow analytic computation of its features. In such cases, one approximates the bootstrap estimate using simulation. First, draw a large number (say, v) of random samples $\boldsymbol{X}^{*(1)}, \ldots, \boldsymbol{X}^{*(v)}$ from the distribution $\hat{F}$ and then compute $T^{(i)} = \eta(\boldsymbol{X}^{*(i)}, \hat{F})$ for $i = 1, \ldots, v$. Finally, compute the desired feature of the sample c.d.f. of the values $T^{(1)}, \ldots, T^{(v)}$.

Example 12.6.3 The M.S.E. of the Sample Median. Suppose that we model our data $\boldsymbol{X} = (X_1, \ldots, X_n)$ as coming from some continuous distribution with the c.d.f. F having median θ. Suppose also that we are interested in using the sample median M as an estimator of θ. We would like to estimate the M.S.E. of M as an estimator of θ. That is, let $\eta(\boldsymbol{X}, F) = (M - \theta)^2$, and try to estimate the mean of $\eta(\boldsymbol{X}, F)$. Let $\hat{F}$ be a known distribution that we hope is similar to F, and let $\boldsymbol{X}^*$ be a random sample of size n from $\hat{F}$. Regardless of what distribution $\hat{F}$ we choose, it is very difficult to compute the bootstrap estimate, the mean of $\eta(\boldsymbol{X}^*, \hat{F})$. Instead, we would simulate a large number v of samples $\boldsymbol{X}^{*(1)}, \ldots, \boldsymbol{X}^{*(v)}$ with the distribution $\hat{F}$ and then compute the sample

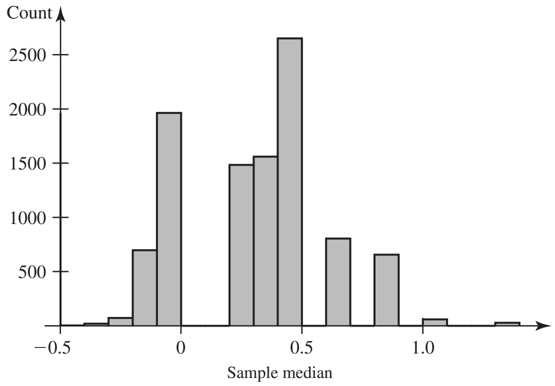


Figure 12.11 Sample medians of 10,000 bootstrap samples in Example 12.6.3.

median of each sample $M^{(1)}, \ldots, M^{(v)}$. Then we compute $T^{(i)} = (M^{(i)} - \hat{\theta})^2$ for $i = 1, \ldots, v$, where $\hat{\theta}$ is the median of the distribution $\hat{F}$. Our simulation approximation to the bootstrap estimate is then the average of the values $T^{(1)}, \ldots, T^{(v)}$.

As an example, suppose that our sample consists of the $n = 25$ values $y_1, \ldots, y_{25}$ listed in Table 10.33 on page 662. For a nonparametric bootstrap analysis, we would use $\hat{F} = F_n$, which is also listed in Table 10.33. Notice that the median of the distribution $\hat{F}$ is the sample median of the original sample, $\hat{\theta} = 0.40$. Next, we simulate v =10,000 random samples of size 25 from the distribution $\hat{F}$. This is done by selecting 25 numbers *with replacement* from the y_i values and repeating for a total of 10,000 samples of size 25. (Solve Exercise 2 to show why this provides the desired samples $\boldsymbol{X}^{*(1)}, \ldots, \boldsymbol{X}^{*(v)}$.) For example, here is one of the 10,000 bootstrap samples:

1.64	0.88	0.70	−1.23	−0.15	1.40	−0.07	−2.46	−2.46	−0.10
−0.15	1.62	0.27	0.44	−0.42	−2.46	1.40	−0.10	0.88	0.44
−1.23	1.07	0.81	−0.02	1.62					

If we sort the numbers in this sample, we find that the sample median is 0.27. In fact, there were 1485 bootstrap samples out of 10,000 that had sample median equal to 0.27. Figure 12.11 contains a histogram of all 10,000 sample medians from the bootstrap samples. The four largest and four smallest observations in the original sample never appeared as sample medians in the 10,000 bootstrap samples. For each of the 10,000 bootstrap samples i, we compute the sample median $M^{(i)}$ and its squared error $T^{(i)} = (M^{(i)} - \hat{\theta})^2$, where $\hat{\theta} = 0.40$ is the median of the distribution $\hat{F}$. We then average all of these values over the 10,000 samples and obtain the value 0.0887. This is our simulation approximation to the nonparametric bootstrap estimate of the M.S.E. of the sample median. The sample variance of the simulated $T^{(i)}$ values is $\hat{\sigma}^2 = 0.0135$, and the simulation standard error of the bootstrap estimate is $\hat{\sigma}/\sqrt{10,000} = 1.163 \times 10^{-3}$. ◀

Note: Simulation Approximation of Bootstrap Estimates. The bootstrap is an estimation technique. As such, it produces estimates of parameters of interest. When a bootstrap estimate is too difficult to compute, we resort to simulation. Simulation provides an estimator of the bootstrap estimate. In this text, we shall refer to the simulation estimator of a bootstrap estimate as an *approximation*. We do this merely to avoid having to refer to estimators of estimates.

The bootstrap was introduced by Efron (1979), and there have been many applications since then. Readers interested in more detail about the bootstrap should see Efron and Tibshirani (1993) or Davison and Hinkley (1997). Young (1994) gives a review of much of the literature on the bootstrap and contains many useful references. In the remainder of this section, we shall present several examples of both the parametric and nonparametric bootstraps and illustrate how simulation is used to approximate the desired bootstrap estimates.

The Nonparametric Bootstrap

Example 12.6.4

Confidence Interval for the Interquartile Range. The interquartile range (IQR) of a distribution was introduced in Definition 4.3.2. It is defined to be the difference between the upper and lower quartiles, the 0.75 and 0.25 quantiles. The central 50 percent of the distribution lies between the lower and upper quartiles, so the IQR is the length of the interval that contains the middle half of the distribution. For example, if F is the normal distribution with variance σ^2, then the IQR is 1.35σ.

Suppose that we desire a 90 percent confidence interval for the IQR θ of the unknown distribution F from which we have a random sample $X_1, \ldots, X_n$. There are many ways to form confidence intervals, so we shall restrict attention to those that are based on the relationship between θ and the sample IQR $\hat{\theta}$. Since the IQR is a scale feature, it might be reasonable to base our confidence interval on the distribution of $\hat{\theta}/\theta$. That is, let the 0.05 and 0.95 quantiles of the distribution of $\hat{\theta}/\theta$ be a and b, so that

$$\Pr\left(a \le \frac{\hat{\theta}}{\theta} \le b\right) = 0.9.$$

Because $a \le \hat{\theta}/\theta \le b$ is equivalent to $\hat{\theta}/b \le \theta \le \hat{\theta}/a$, we conclude that $(\hat{\theta}/b, \hat{\theta}/a)$ is a 90 percent confidence interval for θ. The nonparametric bootstrap can be used to estimate the quantiles a and b as follows: Let $\eta(\boldsymbol{X}, F) = \hat{\theta}/\theta$ be the ratio of the sample IQR of the sample $\boldsymbol{X}$ to the IQR of the distribution F. Let $\hat{F} = F_n$, and notice that the IQR of $\hat{F}$ is $\hat{\theta}$, the sample IQR. Next, let $\boldsymbol{X}^*$ be a sample of size n from $\hat{F}$. Let $\hat{\theta}^*$ be the sample IQR calculated from $\boldsymbol{X}^*$, so that $\eta(\boldsymbol{X}^*, \hat{F}) = \hat{\theta}^*/\hat{\theta}$. The 0.05 and 0.95 quantiles of the distribution of $\eta(\boldsymbol{X}, F)$ are estimated by the 0.05 and 0.95 quantiles of the distribution of $\eta(\boldsymbol{X}^*, \hat{F})$. These last quantiles, in turn, are typically approximated by simulation. We simulate a large number, say, v, of bootstrap samples $\boldsymbol{X}^{*(i)}$ for $i = 1, \ldots, v$. For each bootstrap sample i, we compute the sample IQR $\hat{\theta}^{*(i)}$ and divide it by $\hat{\theta}$. Call the ratio $T^{(i)}$. The q quantile of $\hat{\theta}^*/\hat{\theta}$ is approximated by the sample q quantile of the sample $T^{(1)}, \ldots, T^{(v)}$. The confidence interval constructed by this method is called a *percentile bootstrap confidence interval*.

We can illustrate this with the data in Table 10.33 on page 662. The IQR of the distribution F_n is 1.46, the difference between the 19th and 6th observations. We simulate 10,000 random samples of size 25 from the distribution F_n. For the ith sample, we compute the sample IQR $\hat{\theta}^{*(i)}$ and divide it by 1.46 to obtain $T^{(i)}$. The 500th and 9500th ordered values from $T^{(1)}, \ldots, T^{(10{,}000)}$ are 0.5822 and 1.6301. We then compute the percentile bootstrap confidence interval $(1.46/1.6301, 1.46/0.5822) = (0.8956, 2.5077)$. ◀

Example 12.6.5

Confidence Interval for a Location Parameter. Let $X_1, \ldots, X_n$ be a random sample from the distribution F. Suppose that we want a confidence interval for the median θ of F. We can base a confidence interval on the sample median M. For example,

our interval could be of the form $[M - c_1, M + c_2]$. Since $M - c_1 \le \theta \le M + c_2$ is equivalent to $-c_2 \le M - \theta \le c_1$, we might want $-c_2$ and c_1 to be quantiles of the distribution of $M - \theta$. Without making assumptions about the distribution F, it might be very difficult to approximate quantiles of the distribution of $M - \theta$. To compute a percentile bootstrap confidence interval, let $\eta(\boldsymbol{X}, F) = M - \theta$ and then approximate quantiles (such as $\alpha_0/2$ and $1 - \alpha_0/2$) of the distribution of $\eta(\boldsymbol{X}, F)$ by the corresponding quantiles of $\eta(\boldsymbol{X}^*, \hat{F})$. Here, $\hat{F}$ is the sample c.d.f., F_n, whose median is M, and $\boldsymbol{X}^*$ is a random sample from $\hat{F}$. We then choose a large number v and simulate many samples $\boldsymbol{X}^{*(i)}$ for $i = 1, \ldots, v$. For each sample, we compute the sample median $M^{*(i)}$ and then find the sample quantiles of the values $M^{*(i)} - M$ for $i = 1, \ldots, v$. ◀

How well the percentile bootstrap interval performs in Example 12.6.5 depends on how closely the distribution of $M^* - M$ approximates the distribution of $M - \theta$. (Here, M^* is the median of a sample $\boldsymbol{X}^*$ of size n from $\hat{F}$.) The situation of Example 12.6.5 is one in which there is a possible improvement to the approximation. One thing that can make the distribution of $M^* - M$ different from the distribution of $M - \theta$ is that one of these distributions is more or less spread out than the other. We can use a different bootstrap approximation that suffers less from differences in spread. Instead of constructing an interval of the form $[M - c_1, M + c_2]$, we could let our interval be $[M - d_1Y, M + d_2Y]$, where Y is a statistic that measures the spread of the data. One possibility for Y is the sample IQR. Another possible spread measure is the sample median absolute deviation (the sample median of the values $|X_1 - M|, \ldots, |X_n - M|$). Now, we see that $M - d_1Y \le \theta \le M + d_2Y$ is equivalent to

$$-d_2 \le \frac{M - \theta}{Y} \le d_1.$$

So, we want $-d_2$ and d_1 to be quantiles of the distribution of $(M - \theta)/Y$. This type of interval resembles the t confidence interval developed in Sec. 8.5. Indeed, the interval we are constructing is called a *percentile-t bootstrap confidence interval*. To construct the percentile-t bootstrap confidence interval, we would use each bootstrap sample $\boldsymbol{X}^*$ as follows: Compute the sample median M^* and the scale statistic Y^* from the bootstrap sample $\boldsymbol{X}^*$. Then calculate $T = (M^* - M)/Y^*$. Repeat this procedure many times producing $T^{(1)}, \ldots, T^{(v)}$ from a large number v of bootstrap samples. Then let $-d_2$ and d_1 be sample quantiles (such as $\alpha_0/2$ and $1 - \alpha_0/2$) of the $T^{(i)}$ values.

Example 12.6.6

Percentile-t Confidence Interval for a Median. Consider the $n = 10$ lactic acid concentrations in cheese from Example 8.5.4. We shall do v =10,000 bootstrap simulations to find a coefficient $1 - \alpha_0 = 0.90$ confidence interval for the median lactic acid concentration θ. The median of the sample values is $M = 1.41$, and the median absolute deviation is $Y = 0.245$. The 0.05 and 0.95 sample quantiles of the $(M^{*(i)} - M)/Y^{*(i)}$ values are −2.133 and 1.581. This makes the percentile-t bootstrap confidence interval $(1.41 - 1.581 \times 0.245, 1.41 + 2.133 \times 0.245) = (1.023, 1.933)$. For comparison, the 0.05 and 0.95 sample quantiles of the values of $M^{*(i)} - M$ are −0.32 and 0.16, respectively. This makes the percentile bootstrap interval equal to $(1.41 - 0.16, 1.41 + 0.32) = (1.25, 1.73)$. ◀

The percentile-t interval in Example 11.5.6 is considerably wider than the percentile interval. This reflects the fact that the Y^* values from the bootstrap samples are quite spread out. This in turn suggests that the spread that we should expect to see in a sample has substantial variability. Hence, it is probably not a good idea to assume that the spread of the distribution of $M^* - M$ is the same as the spread of

the distribution of $M - \theta$. The percentile-t bootstrap interval is generally preferred to the percentile bootstrap interval when both are available. This is due to the fact that the distribution of $(M^* - M)/Y^*$ depends less on $\hat{F}$ than does the distribution of $M^* - M$. In particular, $(M^* - M)/Y^*$ does not depend on any scale parameter of the distribution $\hat{F}$. For this reason, we expect more similarity between the distributions of $(M^* - M)/Y^*$ and $(M - \theta)/Y$ than we expect between the distributions of $M^* - M$ and $M - \theta$.

Example 12.6.7

Features of the Distribution of a Sample Correlation. Let (X, Y) have a bivariate joint distribution F with finite variances for both coordinates, so that it makes sense to talk about correlation. Suppose that we observe a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution F. Suppose further that we are interested in the distribution of the sample correlation:

$$R = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\left(\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]\left[\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right]\right)^{1/2}}. \tag{12.6.2}$$

We might be interested in the variance of R, or the bias of R, or some other feature of R as an estimator of the correlation ρ between X and Y. Whatever our goal is, we can make use of the nonparametric bootstrap. For example, consider the bias of R as an estimator of ρ. This bias equals the mean of $\eta(\boldsymbol{X}, F) = R - \rho$. We begin by replacing the joint distribution F by the sample distribution F_n of the observed pairs. This F_n is a discrete joint distribution on pairs of real numbers, and it assigns probability $1/n$ to each of the n observed sample pairs. If (X^*, Y^*) has the distribution F_n, it is easy to check (see Exercise 8) that the correlation between X^* and Y^* is R. We then choose a large number v and simulate v samples of size n from F_n. For each i, we compute the sample correlation $R^{(i)}$ by plugging the ith bootstrap sample into Eq. (12.6.2). For each i, we compute $T^{(i)} = R^{(i)} - R$, and we estimate the mean of $R - \rho$ by the average $\frac{1}{v}\sum_{i=1}^{v} T^{(i)}$.

As a numerical example, consider the flea beetle data from Example 5.10.2. The sample correlation is $R = 0.6401$. We sample $v = 10{,}000$ bootstrap samples of size $n = 31$. The average sample correlation in the 10,000 bootstrap samples is 0.6354 with a simulation standard error of 0.001. We then estimate the bias of the sample correlation to be $0.6354 - 0.6401 = -0.0047$. ◀

The Parametric Bootstrap

Example 12.6.8

Correcting the Bias in the Coefficient of Variation. The coefficient of variation of a distribution is the ratio of the standard deviation to the mean. (Typically, people only compute the coefficient of variation for distributions of positive random variables.) If we believe that our data $X_1, \ldots, X_n$ come from a lognormal distribution with parameters μ and σ^2, then the coefficient of variation is $\theta = (e^{\sigma^2} - 1)^{1/2}$. The M.L.E. of the coefficient of variation is $\hat{\theta} = (e^{\hat{\sigma}^2} - 1)^{1/2}$, where $\hat{\sigma}$ is the M.L.E. of σ. We expect the M.L.E. of the coefficient of variation to be a biased estimator because it so nonlinear. Computing the bias is a difficult task. However, we can use the parametric bootstrap to estimate the bias. The M.L.E. $\hat{\sigma}$ of σ is the square root of the sample variance of $\log(X_1), \ldots, \log(X_n)$. The M.L.E. $\hat{\mu}$ of μ is the sample average of $\log(X_1), \ldots, \log(X_n)$. We can simulate a large number of random samples of size n from the lognormal distribution with parameters $\hat{\mu}$ and $\hat{\sigma}^2$. For each i, we compute

$\hat{\sigma}^{*(i)}$, the sample standard deviation of the ith bootstrap sample. We estimate the bias of $\hat{\theta}$ by the sample average of the values $T^{(i)} = (e^{[\hat{\sigma}^{*(i)}]^2} - 1)^{1/2} - \hat{\theta}$.

As an example, consider the failure times of ball bearings introduced in Example 5.6.9. If we model these data as lognormal, the M.L.E.'s of the parameters are $\hat{\mu} = 4.150$ and $\hat{\sigma} = 0.5217$. The M.L.E. of θ is $\hat{\theta} = 0.5593$. We could draw 10,000 random samples of size 23 from a lognormal distribution and compute the sample variances of the logarithms. However, there is an easier way to do this simulation. The distribution of $[\hat{\sigma}^{*(i)}]^2$ is that of a χ^2 random variable with 22 degrees of freedom times $0.5217^2/23$. Hence, we shall just sample 10,000 χ^2 random variables with 22 degrees of freedom, multiply each one by $0.5217^2/23$, and call the ith one $[\hat{\sigma}^{*(i)}]^2$. After doing this, the sample average of the 10,000 $T^{(i)}$ values is -0.01825, which is our parametric bootstrap estimate of the bias of $\hat{\theta}$. (The simulation standard error is 9.47×10^{-4}.) Because our estimate of the bias is negative, this means that we expect $\hat{\theta}$ to be smaller than θ. To "correct" the bias, we could add 0.01825 to our original estimate $\hat{\theta}$ and produce the new estimate $0.5593 + 0.01825 = 0.5776$. ◀

Example 12.6.9 Estimating the Standard Deviation of a Statistic. Suppose that $X_1, \ldots, X_n$ is a random sample from the normal distribution with mean μ and variance σ^2. We are interested in the probability that a random variable having this same distribution is at most c. That is, we are interested in estimating $\theta = \Phi([c - \mu]/\sigma)$. The M.L.E. of θ is $\hat{\theta} = \Phi([c - \overline{X}]/\hat{\sigma})$. It is not easy to calculate the standard deviation of $\hat{\theta}$ in closed form. However, we can draw many, say, v, bootstrap samples of size n from the normal distribution with mean $\overline{x}$ and variance $\hat{\sigma}^2$. For the ith bootstrap sample, we compute a sample average $\overline{x}^{*(i)}$, a sample standard deviation $\hat{\sigma}^{*(i)}$, and, finally, $\hat{\theta}^{*(i)} = \Phi([c - \overline{x}^{*(i)}]/\hat{\sigma}^{*(i)})$. We estimate the mean of $\hat{\theta}$ by

$$\overline{\theta}^* = \frac{1}{v}\sum_{i=1}^{v} \hat{\theta}^{*(i)}.$$

(This can also be used, as in Example 12.6.8, to estimate the bias of $\hat{\theta}$.) The standard deviation of $\hat{\theta}$ can then be estimated by the sample standard deviation of the $\hat{\theta}^{*(i)}$ values,

$$Z = \left(\frac{1}{v}\sum_{i=1}^{v}(\hat{\theta}^{*(i)} - \overline{\theta}^*)^2\right)^{1/2}.$$

For example, we can use the nursing home data from Sec. 8.6. There are $n = 18$ observations, and we might be interested in $\Phi([200 - \mu]/\sigma)$. The M.L.E.'s of μ and σ are $\hat{\mu} = 182.17$ and $\hat{\sigma} = 72.22$. The observed value of $\hat{\theta}$ is $\Phi([200 - 182.17]/72.22) = 0.5975$. We simulate 10,000 samples of size 18 from the normal distribution with mean 182.17 and variance $(72.22)^2$. For the ith sample, we find the value $\hat{\theta}^{*(i)}$ for $i = 1, \ldots, 10{,}000$, and the average of these is $\overline{\theta}^* = 0.6020$ with sample standard deviation $Z = 0.09768$.

We can compute the simulation standard error of the approximation to the bootstrap estimate in two steps. First, apply the method of Example 12.2.10. This gives the simulation standard error of Z^2, the sample variance of the $\hat{\theta}^{*(i)}$'s. In our example, this yields the value 1.365×10^{-4}. Second, use the delta method, as in Example 12.2.8, to find the simulation standard error of the square root of Z^2. In our example, this second step yields the value 6.986×10^{-4}. ◀

Example 12.6.10 Comparing Means When Variances Are Unequal. Suppose that we have two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from two possibly different normal distributions. That is, $X_1, \ldots, X_m$ are i.i.d. from the normal distribution with mean μ_1 and variance σ_1^2, while $Y_1, \ldots, Y_n$ are i.i.d. from the normal distribution with mean μ_2 and variance σ_2^2. In Sec. 9.6, we saw how to test the null hypothesis $H_0 : \mu_1 = \mu_2$ versus the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ if we are willing to assume that we know the ratio $k = \sigma_2^2/\sigma_1^2$. If we are not willing to assume that we know the ratio k, we have seen only approximate tests.

Suppose that we choose to use the usual two-sample t test even though we do not claim to know k. That is, suppose that we choose to reject H_0 when $|U| > c$, where U is the statistic defined in Eq. (9.6.3) and c is the $1 - \alpha_0/2$ quantile of the t distribution with $m + n - 2$ degrees of freedom. This test will not necessarily have level α_0 if $k \neq 1$. We can use the parametric bootstrap to try to compute the level of this test. In fact, we can use the parametric bootstrap to help us choose a different critical value c^* for the test so that we at least estimate the type I error probability to be α_0.

As an example, we shall use the data from Example 9.6.5 again. The M.L.E.'s of the variances of the two distributions were $\hat{\sigma}_1^2 = 0.04$ (for the X data) and $\hat{\sigma}_2^2 = 0.022$ (for the Y data). The probability of type I error is the probability of rejecting the null hypothesis given that the null hypothesis is true, that is, given that $\mu_1 = \mu_2$. Hence, we must simulate bootstrap samples in which the X data and Y data have the same mean. Since the sample averages of the X and Y data are subtracted from each other in the calculation of U, it will not matter what common mean we choose for the two samples.

So, the parametric bootstrap can proceed as follows: First, choose a large number v, and for $i = 1, \ldots, v$, simulate $(\overline{X}^{*(i)}, \overline{Y}^{*(i)}, S_X^{2*(i)}, S_Y^{2*(i)})$ where all four random variables are independent with the following distributions:

- $\overline{X}^{*(i)}$ has the normal distribution with mean 0 and variance $\hat{\sigma}_1^2/m$.
- $\overline{Y}^{*(i)}$ has the normal distribution with mean 0 and variance $\hat{\sigma}_2^2/n$.
- $S_X^{2*(i)}$ is $\hat{\sigma}_1^2$ times a random variable having the χ^2 distribution with $m - 1$ degrees of freedom.
- $S_Y^{2*(i)}$ is $\hat{\sigma}_2^2$ times a random variable having the χ^2 distribution with $n - 1$ degrees of freedom.

Then compute

$$U^{(i)} = \frac{(m + n - 2)^{1/2}(\overline{X}^{*(i)} - \overline{Y}^{*(i)})}{\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}\left(S_X^{2*(i)} + S_Y^{2*(i)}\right)^{1/2}}$$

for each i. Our simulation approximation to the bootstrap estimate of the probability of type I error for the usual two-sample t test would be the proportion of simulations in which $|U^{(i)}| > c$.

With $v = 10{,}000$, we shall perform the analysis described above for several different c values. We set c equal to the $1 - \alpha_0/2$ quantile of the t distribution with 16 degrees of freedom with $\alpha_0 = j/1000$ for each $j = 1, \ldots, 999$. Figure 12.12 shows a plot of the simulation approximation to the bootstrap estimate of the type I error probability against the nominal level α_0 of the test. There is remarkably close agreement between the two, although the bootstrap estimate is generally slightly larger. For example, when $\alpha_0 = 0.05$, the bootstrap estimate is 0.065.

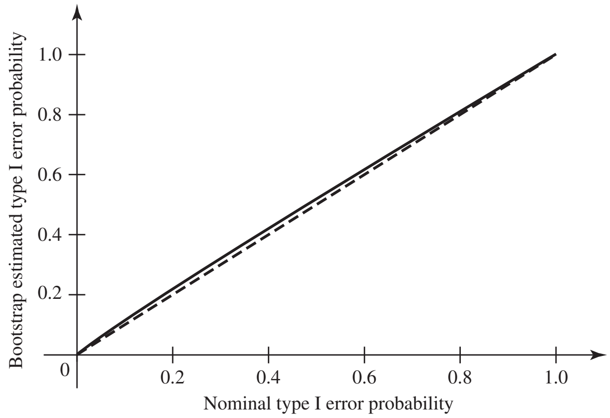


Figure 12.12 Plots of bootstrap estimated type I error probability of t test versus nominal type I error probability in Example 12.6.10. The dashed line is the diagonal along which the two error probabilities would be equal.

Next, we use the bootstrap analysis to correct the level of the two-sample t test in this example. To do this, let Z be the sample $1-\alpha_0$ quantile of our simulated $|U^{(i)}|$ values. If we want a level α_0 test, we can replace the critical value c in the two-sample t test with Z and reject the null hypothesis if $|U| > Z$. For example, with $\alpha_0 = 0.05$, the 0.975 quantile of the t distribution is 2.12, while in our simulation $Z = 2.277$. The simulation standard error of Z (based on splitting the 10,000 bootstrap samples into 10 subsamples of 1000 each) is 0.0089. ◀

Example 12.6.11

The Bias of the Sample Correlation. In Example 12.6.7, we made no assumptions about the distribution F of (X, Y) except that X and Y have finite variances. Now suppose that we also assume that (X, Y) has a bivariate normal distribution. We can compute the M.L.E.'s of all of the parameters as in Exercise 24 in Sec. 7.6. We could then simulate v samples of size n from the bivariate normal distribution with parameters equal to the M.L.E.'s, as in Example 12.3.6. For sample i for $i = 1, \dots, v$, we could compute the sample correlation $R^{(i)}$ by substituting the ith sample into Eq. (12.6.2). Our estimate of the bias would be $\overline{R} - \hat{\rho}$. Note that $\hat{\rho}$, the M.L.E. of ρ, is the same as R.

As a numerical example, consider the flea beetle data from Example 5.10.2. The sample correlation is $R = 0.6401$. We construct $v = 10{,}000$ samples of size $n = 31$ from a bivariate normal distribution with correlation 0.6401. The means and variances do not affect the distribution of R. (See Exercise 12.) The average sample correlation in the 10,000 bootstrap samples is 0.6352 with a simulation standard error of 0.001. We then estimate the bias of the sample correlation to be $0.6352 - 0.6401 = -0.0049$. This is pretty much the same as we obtained using the nonparametric bootstrap in Example 12.6.7. ◀

Summary

The bootstrap is a method for estimating probabilistic features of a function η of our data $\boldsymbol{X}$ and their unknown distribution F. That is, suppose that we are interested in the mean, a quantile, or some other feature of $\eta(\boldsymbol{X}, F)$. The first step in the bootstrap is to replace F by a known distribution $\hat{F}$ that is like F in some way. Next, replace $\boldsymbol{X}$ by data $\boldsymbol{X}^*$ sampled from $\hat{F}$. Finally, compute the mean, quantile, or other feature of $\eta(\boldsymbol{X}^*, \hat{F})$ as the bootstrap estimate. This last step generally requires simulation except in the simplest examples. There are two varieties of bootstrap that differ by

how $\hat{F}$ is chosen. In the nonparametric bootstrap, the sample c.d.f. is used as $\hat{F}$. In the parametric bootstrap, F is assumed to be a member of some parametric family and $\hat{F}$ is chosen by replacing the unknown parameter by its M.L.E. or some other estimate.

Exercises

1. Suppose that $X_1, \ldots, X_n$ form a random sample from an exponential distribution with parameter θ. Explain how to use the parametric bootstrap to estimate the variance of the sample average $\overline{X}$. (No simulation is required.)

2. Let $x_1, \ldots, x_n$ be the observed values of a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$. Let F_n be the sample c.d.f. Let $J_1, \ldots, J_n$ be a random sample with replacement from the numbers $\{1, \ldots, n\}$. Define $X_i^* = x_{J_i}$ for $i = 1, \ldots, n$. Show that $\boldsymbol{X}^* = (X_1^*, \ldots, X_n^*)$ is an i.i.d. sample from the distribution F_n.

3. Let n be odd, and let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sample of size n from some distribution. Suppose that we wish to use the nonparametric bootstrap to estimate some feature of the sample median. Compute the probability that the sample median of a nonparametric bootstrap sample will be the smallest observation from the original data $\boldsymbol{X}$.

4. Use the data in the first column of Table 11.5 on page 699. These data give the boiling points of water at 17 different locations from Forbes' experiment. Let F be the distribution from which these boiling points were drawn. We might not be willing to make many assumptions about F. Suppose that we are interested in the bias of the sample median as an estimator of the median of the distribution F. Use the nonparametric bootstrap to estimate this bias. First, do a pilot run to compute the simulation standard error of the simulation approximation, and then see how many bootstrap samples you need in order for your bias estimate (for distribution $\hat{F}$) to be within 0.02 of the true bias (for distribution $\hat{F}$) with probability at least 0.9.

5. Use the data in Table 10.6 on page 640. We are interested in the bias of the sample median as an estimator of the median of the distribution.

- **a.** Use the nonparametric bootstrap to estimate this bias.
- **b.** How many bootstrap samples does it appear that you need in order to estimate the bias to within .05 with probability 0.99?

6. Use the data in Exercise 16 of Sec. 10.7.

- **a.** Use the nonparametric bootstrap to estimate the variance of the sample median.
- **b.** How many bootstrap samples does it appear that you need in order to estimate the variance to within .005 with probability 0.95?

7. Use the blood pressure data in Table 9.2 that was described in Exercise 10 of Sec. 9.6. Suppose now that we are not confident that the variances are the same for the two treatment groups. Perform a parametric bootstrap analysis of the sort done in Example 12.6.10. Use $v =$10,000 bootstrap simulations.

- **a.** Estimate the probability of type I error for a two-sample t test whose nominal level is $\alpha_0 = 0.1$.
- **b.** Correct the level of the two-sample t test by computing the appropriate quantile of the bootstrap distribution of $|U^{(i)}|$.
- **c.** Compute the simulation standard error for the quantile in part (b).

8. In Example 12.6.7, let (X^*, Y^*) be a random draw from the sample distribution F_n. Prove that the correlation between X^* and Y^* is R in Eq. (12.6.2).

9. Use the data on fish prices in Table 11.6 on page 707. Suppose that we assume only that the distribution of fish prices in 1970 and 1980 is a continuous joint distribution with finite variances. We are interested in the properties of the sample correlation coefficient. Construct 1000 nonparametric bootstrap samples for solving this exercise.

- **a.** Approximate the bootstrap estimate of the variance of the sample correlation.
- **b.** Approximate the bootstrap estimate of the bias of the sample correlation.
- **c.** Compute simulation standard errors of each of the above bootstrap estimates.

10. Use the beef hot dog data in Exercise 7 of Sec. 8.5. Form 10,000 nonparametric bootstrap samples to solve this exercise.

- **a.** Approximate a 90 percent percentile bootstrap confidence interval for the median calorie count in beef hot dogs.
- **b.** Approximate a 90 percent percentile-t bootstrap confidence interval for the median calorie count in beef hot dogs.
- **c.** Compare these intervals to the 90 percent interval formed using the assumption that the data came from a normal distribution.

11. The skewness of a random variable was defined in Definition 4.4.1. Suppose that $X_1, \dots, X_n$ form a random sample from a distribution F. The sample skewness is defined as

$$M_3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{3/2}}.$$

One might use M_3 as an estimator of the skewness θ of the distribution F. The bootstrap can be used to estimate the bias and standard deviation of the sample skewness as an estimator of θ.

a. Prove that M_3 is the skewness of the sample distribution F_n.

b. Use the 1970 fish price data in Table 11.6 on page 707. Compute the sample skewness, and then simulate 1000 bootstrap samples. Use the bootstrap samples to estimate the bias and standard deviation of the sample skewness.

12. Suppose that $(X_1, Y_1), \dots, (X_n, Y_n)$ form a random sample from a bivariate normal distribution with means μ_x and μ_y, variances σ_x^2 and σ_y^2, and correlation ρ. Let R be the sample correlation. Prove that the distribution of R depends only on ρ, not on μ_x, μ_y, σ_x^2, or σ_y^2.

12.7 Supplementary Exercises

1. Test the standard normal pseudo-random number generator on your computer by generating a sample of size 10,000 and drawing a normal quantile plot. How straight does the plot appear to be?

2. Test the gamma pseudo-random number generator on your computer. Simulate 10,000 gamma pseudo-random variables with parameters a and 1 for $a = 0.5, 1, 1.5, 2, 5, 10$. Then draw gamma quantile plots.

3. Test the t pseudo-random number generator on your computer. Simulate 10,000 t pseudo-random variables with m degrees of freedom for $m = 1, 2, 5, 10, 20$. Then draw t quantile plots.

4. Let X and Y be independent random variables with X having the t distribution with five degrees of freedom and Y having the t distribution with three degrees of freedom. We are interested in $E(|X - Y|)$.

a. Simulate 1000 pairs of (X_i, Y_i) each with the above joint distribution and estimate $E(|X - Y|)$.

b. Use your 1000 simulated pairs to estimate the variance of $|X - Y|$ also.

c. Based on your estimated variance, how many simulations would you need to be 99 percent confident that your estimator of $E(|X - Y|)$ is within 0.01 of the actual mean?

5. Consider the power calculation done in Example 9.5.5.

a. Simulate $v_0 = 1000$ i.i.d. noncentral t pseudo-random variables with 14 degrees of freedom and noncentrality parameter 1.936.

b. Estimate the probability that a noncentral t random variable with 14 degrees of freedom and noncentrality parameter 1.936 is at least 1.761. Also, compute the simulation standard error.

c. Suppose that we want our estimator of the noncentral t probability in part (b) to be closer than 0.01 to the true value with probability 0.99. How many noncentral t random variables do we need to simulate?

6. The χ^2 goodness-of-fit test (see Chapter 10) is based on an asymptotic approximation to the distribution of the test statistic. For small to medium samples, the asymptotic approximation might not be very good. Simulation can be used to assess how good the approximation is. Simulation can also be used to estimate the power function of a goodness-of-fit test. For this exercise, assume that we are performing the test that was done in Example 10.1.6. The idea illustrated in this exercise applies in all such problems.

a. Simulate $v = 10{,}000$ samples of size $n = 23$ from the normal distribution with mean 3.912 and variance 0.25. For each sample, compute the χ^2 goodness-of-fit statistic Q using the same four intervals that were used in Example 10.1.6. Use the simulations to estimate the probability that Q is greater than or equal to the 0.9, 0.95, and 0.99 quantiles of the χ^2 distribution with three degrees of freedom.

b. Suppose that we are interested in the power function of a χ^2 goodness-of-fit test when the actual distribution of the data is the normal distribution with mean 4.2 and variance 0.8. Use simulation to estimate the power function of the level 0.1, 0.05, and 0.01 tests at the alternative specified.

7. In Sec. 10.2, we discussed χ^2 goodness-of-fit tests for composite hypotheses. These tests required computing M.L.E.'s based on the numbers of observations that fell into the different intervals used for the test. Suppose instead that we use the M.L.E.'s based on the original observations. In this case, we claimed that the asymptotic distribution of the χ^2 test statistic was somewhere between two different χ^2 distributions. We can use simulation to better approximate the distribution of the test statistic. In this exercise, assume that we are trying to test

the same hypotheses as in Example 10.2.5, although the methods will apply in all such cases.

a. Simulate $v = 1000$ samples of size $n = 23$ from each of 10 different normal distributions. Let the normal distributions have means of 3.8, 3.9, 4.0, 4.1, and 4.2. Let the distributions have variances of 0.25 and 0.8. Use all 10 combinations of mean and variance. For each simulated sample, compute the χ^2 statistic Q using the usual M.L.E.'s of μ and σ^2. For each of the 10 normal distributions, estimate the 0.9, 0.95, and 0.99 quantiles of the distribution of Q.

b. Do the quantiles change much as the distribution of the data changes?

c. Consider the test that rejects the null hypothesis if $Q \geq 5.2$. Use simulation to estimate the power function of this test at the following alternative: For each i, $(X_i - 3.912)/0.5$ has the t distribution with five degrees of freedom.

8. In Example 12.5.6, we used a hierarchical model. In that model, the parameters $\mu_1, \dots, \mu_p$ were independent random variables with μ_i having the normal distribution with mean ψ and precision $\lambda_0\tau_i$ conditional on ψ and $\tau_1, \dots, \tau_p$. To make the model more general, we could also replace λ_0 by an unknown parameter λ. That is, let the μ_i's be independent with μ_i having the normal distribution with mean ψ and precision $\lambda\tau_i$ conditional on ψ, λ, and $\tau_1, \dots, \tau_p$. Let λ have the gamma distribution with parameters γ_0 and δ_0, and let λ be independent of ψ and $\tau_1, \dots, \tau_p$. The remaining parameters have the prior distributions stated in Example 12.5.6.

a. Write the product of the likelihood and the prior as a function of the parameters $\mu_1, \dots, \mu_p, \tau_1, \dots, \tau_p$, ψ, and λ.

b. Find the conditional distributions of each parameter given all of the others. *Hint:* For all the parameters besides λ, the distributions should be almost identical to those given in Example 12.5.6. Wherever λ_0 appears, of course, something will have to change.

c. Use a prior distribution in which $\alpha_0 = 1$, $\beta_0 = 0.1$, $u_0 = 0.001$, $\gamma_0 = \delta_0 = 1$, and $\psi_0 = 170$. Fit the model to the hot dog calorie data from Example 11.6.2. Compute the posterior means of the four μ_i's and $1/\tau_i$'s.

9. In Example 12.5.6, we modeled the parameters $\tau_1, \dots, \tau_p$ as i.i.d. having the gamma distribution with parameters α_0 and β_0. We could have added a level to the hierarchical model that would allow the τ_i's to come from a distribution with an unknown parameter. For example, suppose that we model the τ_i's as conditionally independent having the gamma distribution with parameters α_0 and β given β. Let β be independent of ψ and $\mu_1, \dots, \mu_p$ with β having the gamma distribution with parameters ϵ_0 and ϕ_0. The rest of the prior distributions are as specified in Example 12.5.6.

a. Write the product of the likelihood and the prior as a function of the parameters $\mu_1, \dots, \mu_p, \tau_1, \dots, \tau_p$, ψ, and β.

b. Find the conditional distributions of each parameter given all of the others. *Hint:* For all the parameters besides β, the distributions should be almost identical to those given in Example 12.5.6. Wherever β_0 appears, of course, something will have to change.

c. Use a prior distribution in which $\alpha_0 = \lambda_0 = 1$, $u_0 = 0.001$, $\epsilon_0 = 0.3$, $\phi_0 = 3.0$, and $\psi_0 = 170$. Fit the model to the hot dog calorie data from Example 11.6.2. Compute the posterior means of the four μ_i's and $1/\tau_i$'s.

10. Let $X_1, \dots, X_k$ be independent random variables such that X_i has the binomial distribution with parameters n_i and p_i. We wish to test the null hypothesis H_0: $p_1 = \cdots = p_k$ versus the alternative hypothesis H_1 that H_0 is false. Assume that the numbers $n_1, \dots, n_k$ are known constants.

a. Show that the likelihood ratio test procedure is to reject H_0 if the following statistic is greater than or equal to some constant c:

$$\frac{\prod_{i=1}^{k}\left[X_i^{X_i}(n_i - X_i)^{n_i - X_i}\right]}{\left(\sum_{j=1}^{k} X_j\right)^{\sum_{j=1}^{k} X_j}\left[\sum_{j=1}^{k}(n_j - X_j)\right]^{\sum_{j=1}^{k}(n_j - X_j)}}.$$

b. Describe how you could use simulation techniques to estimate the constant c in order to make the likelihood ratio test have a desired level of significance α_0. (Assume that you can simulate as many binomial pseudo-random variables as you wish.)

c. Consider the depression study in Example 2.1.4. Let p_i stand for the probability of success (no relapse) for the subjects in group i of Table 2.1 on page 57, where $i = 1$ means imipramine, $i = 2$ means lithium, $i = 3$ means combination, and $i = 4$ means placebo. Test the null hypothesis that $p_1 = p_2 = p_3 = p_4$ by computing the p-value for the likelihood ratio test.

11. Consider the problem of testing the equality of two normal means when the variances are unequal. This problem was introduced on page 593 in Sec. 9.6. The data are two independent samples $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$. The X_i's are i.i.d. having the normal distribution with mean μ_1 and variance σ_1^2, while the Y_j's are i.i.d. having the normal distribution with mean μ_2 and variance σ_2^2.

a. Assume that $\mu_1 = \mu_2$. Prove that the random variable V in Eq. (9.6.14) has a distribution that depends on the parameters only through the ratio σ_2/σ_1.

b. Let ν be the approximate degrees of freedom for Welch's procedure from Eq. (9.6.17). Prove that the distribution of ν depends on the parameters only through the ratio σ_2/σ_1.

c. Use simulation to assess the approximation in Welch's procedure. In particular, set the ratio σ_2/σ_1 equal to each of the numbers 1, 1.5, 2, 3, 5, and 10 in succession. For each value of the ratio, simulate 10,000 samples of sizes $n = 11$ and $m = 10$ (or the appropriate summary statistics). For each simulated sample, compute the test statistic V and the 0.9, 0.95, and 0.99 quantiles of the approximate t distribution that corresponds to the data in that simulation. Keep track of the proportion of simulations in which V is greater than each of the three quantiles. How do these proportions compare to the nominal values 0.1, 0.05, and 0.01?

12. Consider again the situation described in Exercise 11. This time, use simulation to assess the performance of the usual two-sample t test. That is, use the same simulations as in part (c) of Exercise 11 (or ones just like them if you do not have the same simulations). This time, for each simulated sample compute the statistic U in Eq. (9.6.3) and keep track of the proportion of simulations in which U is greater than each of the nominal t quantiles, $T_{19}^{-1}(1-\alpha_0)$ for $\alpha_0 = 0.1$, 0.05, and 0.01. How do these proportions compare to the nominal α_0 values?

13. Suppose that our data comprise a set of pairs (Y_i, x_i), for $i = 1, \ldots, n$. Here, each Y_i is a random variable and each x_i is a known constant. Suppose that we use a simple linear regression model in which $E(Y_i) = \beta_0 + \beta_1 x_i$. Let $\hat{\beta}_1$ stand for the least squares estimator of β_1. Suppose, however, that the Y_i's are actually random variables with translated and scaled t distributions. In particular, suppose that $(Y_i - \beta_0 - \beta_1 x_i)/\sigma$ are i.i.d. having the t distribution with $k \ge 5$ degrees of freedom for $i = 1, \ldots, n$. We can use simulation to estimate the standard deviation of the sampling distribution of $\hat{\beta}_1$.

a. Prove that the variance of the sampling distribution of $\hat{\beta}_1$ does not depend on the values of the parameters β_0 and β_1.

b. Prove that the variance of the sampling distribution of $\hat{\beta}_1$ is equal to $v\sigma^2$, where v does not depend on any of the parameters β_0, β_1, and σ.

c. Describe a simulation scheme to estimate the value v from part (b).

14. Use the simulation scheme developed in Exercise 13 and the data in Table 11.5 on page 699. Suppose that we think that the logarithms of pressure are linearly related to boiling point, but that the logarithms of pressure have translated and scaled t distributions with $k = 5$ degrees of freedom. Estimate the value v from part (b) of Exercise 13 using simulation.

15. In Sec. 7.4, we introduced Bayes estimators. For simple loss functions, such as squared error and absolute error, we were able to derive general forms for Bayes estimators. In many real problems, loss functions are not so simple. Simulation can often be used to approximate Bayes estimators. Suppose that we are able to simulate a sample $\theta^{(1)}, \ldots, \theta^{(v)}$ (either directly or by Gibbs sampling) from the posterior distribution of some parameter θ given some observed data $X = x$. Here, θ can be real valued or multidimensional. Suppose that we have a loss function $L(\theta, a)$, and we want to choose a so as to minimize the posterior mean $E[L(\theta, a)|\boldsymbol{x}]$.

a. Describe a general method for approximating the Bayes estimate in the situation described above.

b. Suppose that the simulation variance of the approximation to the Bayes estimate is proportional to 1 over the size of the simulation. How could one compute a simulation standard error for the approximation to the Bayes estimate?

16. In Example 12.5.2, suppose that the State of New Mexico wishes to estimate the mean number μ of medical in-patient days in nonrural nursing homes. The parameter is $\theta = (\mu, \tau)$. The loss function will be asymmetric to reflect different costs of underestimating and overestimating. Suppose that the loss function is

$$L(\theta, a) = \begin{cases} 30(a-\mu) & \text{if } a \ge \mu, \\ (\mu - a)^2 & \text{if } \mu > a. \end{cases}$$

Use the method developed in your solution to Exercise 15 to approximate the Bayes estimate and compute a simulation standard error.

Tables

Table of Binomial Probabilities

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

n	k	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$
2	0	.8100	.6400	.4900	.3600	.2500
	1	.1800	.3200	.4200	.4800	.5000
	2	.0100	.0400	.0900	.1600	.2500
3	0	.7290	.5120	.3430	.2160	.1250
	1	.2430	.3840	.4410	.4320	.3750
	2	.0270	.0960	.1890	.2880	.3750
	3	.0010	.0080	.0270	.0640	.1250
4	0	.6561	.4096	.2401	.1296	.0625
	1	.2916	.4096	.4116	.3456	.2500
	2	.0486	.1536	.2646	.3456	.3750
	3	.0036	.0256	.0756	.1536	.2500
	4	.0001	.0016	.0081	.0256	.0625
5	0	.5905	.3277	.1681	.0778	.0312
	1	.3280	.4096	.3602	.2592	.1562
	2	.0729	.2048	.3087	.3456	.3125
	3	.0081	.0512	.1323	.2304	.3125
	4	.0005	.0064	.0284	.0768	.1562
	5	.0000	.0003	.0024	.0102	.0312
6	0	.5314	.2621	.1176	.0467	.0156
	1	.3543	.3932	.3025	.1866	.0938
	2	.0984	.2458	.3241	.3110	.2344
	3	.0146	.0819	.1852	.2765	.3125
	4	.0012	.0154	.0595	.1382	.2344
	5	.0001	.0015	.0102	.0369	.0938
	6	.0000	.0001	.0007	.0041	.0156
7	0	.4783	.2097	.0824	.0280	.0078
	1	.3720	.3670	.2471	.1306	.0547
	2	.1240	.2753	.3176	.2613	.1641
	3	.0230	.1147	.2269	.2903	.2734
	4	.0026	.0287	.0972	.1935	.2734
	5	.0002	.0043	.0250	.0774	.1641
	6	.0000	.0004	.0036	.0172	.0547
	7	.0000	.0000	.0002	.0016	.0078

(continued)

Table of Binomial Probabilities *(continued)*

n	k	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$
8	0	.4305	.1678	.0576	.0168	.0039
	1	.3826	.3355	.1977	.0896	.0312
	2	.1488	.2936	.2965	.2090	.1094
	3	.0331	.1468	.2541	.2787	.2188
	4	.0046	.0459	.1361	.2322	.2734
	5	.0004	.0092	.0467	.1239	.2188
	6	.0000	.0011	.0100	.0413	.1094
	7	.0000	.0001	.0012	.0079	.0312
	8	.0000	.0000	.0001	.0007	.0039
9	0	.3874	.1342	.0404	.0101	.0020
	1	.3874	.3020	.1556	.0605	.0176
	2	.1722	.3020	.2668	.1612	.0703
	3	.0446	.1762	.2668	.2508	.1641
	4	.0074	.0661	.1715	.2508	.2461
	5	.0008	.0165	.0735	.1672	.2461
	6	.0001	.0028	.0210	.0743	.1641
	7	.0000	.0003	.0039	.0212	.0703
	8	.0000	.0000	.0004	.0035	.0176
	9	.0000	.0000	.0000	.0003	.0020
10	0	.3487	.1074	.0282	.0060	.0010
	1	.3874	.2684	.1211	.0403	.0098
	2	.1937	.3020	.2335	.1209	.0439
	3	.0574	.2013	.2668	.2150	.1172
	4	.0112	.0881	.2001	.2508	.2051
	5	.0015	.0264	.1029	.2007	.2461
	6	.0001	.0055	.0368	.1115	.2051
	7	.0000	.0008	.0090	.0425	.1172
	8	.0000	.0001	.0014	.0106	.0439
	9	.0000	.0000	.0001	.0016	.0098
	10	.0000	.0000	.0000	.0001	.0010

(continued)

Table of Binomial Probabilities *(continued)*

n	k	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$
15	0	.2059	.0352	.0047	.0005	.0000
	1	.3432	.1319	.0305	.0047	.0005
	2	.2669	.2309	.0916	.0219	.0032
	3	.1285	.2501	.1700	.0634	.0139
	4	.0428	.1876	.2186	.1268	.0417
	5	.0105	.1032	.2061	.1859	.0916
	6	.0019	.0430	.1472	.2066	.1527
	7	.0003	.0138	.0811	.1771	.1964
	8	.0000	.0035	.0348	.1181	.1964
	9	.0000	.0007	.0116	.0612	.1527
	10	.0000	.0001	.0030	.0245	.0916
	11	.0000	.0000	.0006	.0074	.0417
	12	.0000	.0000	.0001	.0016	.0139
	13	.0000	.0000	.0000	.0003	.0032
	14	.0000	.0000	.0000	.0000	.0005
	15	.0000	.0000	.0000	.0000	.0000
20	0	.1216	.0115	.0008	.0000	.0000
	1	.2701	.0576	.0068	.0005	.0000
	2	.2852	.1369	.0278	.0031	.0002
	3	.1901	.2054	.0716	.0123	.0011
	4	.0898	.2182	.1304	.0350	.0046
	5	.0319	.1746	.1789	.0746	.0148
	6	.0089	.1091	.1916	.1244	.0370
	7	.0020	.0545	.1643	.1659	.0739
	8	.0003	.0222	.1144	.1797	.1201
	9	.0001	.0074	.0654	.1597	.1602
	10	.0000	.0020	.0308	.1171	.1762
	11	.0000	.0005	.0120	.0710	.1602
	12	.0000	.0001	.0039	.0355	.1201
	13	.0000	.0000	.0010	.0146	.0739
	14	.0000	.0000	.0002	.0049	.0370
	15	.0000	.0000	.0000	.0013	.0148
	16	.0000	.0000	.0000	.0003	.0046
	17	.0000	.0000	.0000	.0000	.0011
	18	.0000	.0000	.0000	.0000	.0002
	19	.0000	.0000	.0000	.0000	.0000
	20	.0000	.0000	.0000	.0000	.0000

Table of Poisson Probabilities

$$\Pr(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

k	$\lambda = .1$	.2	.3	.4	.5	.6	.7	.8	.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

k	$\lambda = 1.5$	2	3	4	5	6	7	8	9	10
0	.2231	.1353	.0498	.0183	.0067	.0025	.0009	.0003	.0001	.0000
1	.3347	.2707	.1494	.0733	.0337	.0149	.0064	.0027	.0011	.0005
2	.2510	.2707	.2240	.1465	.0842	.0446	.0223	.0107	.0050	.0023
3	.1255	.1804	.2240	.1954	.1404	.0892	.0521	.0286	.0150	.0076
4	.0471	.0902	.1680	.1954	.1755	.1339	.0912	.0573	.0337	.0189
5	.0141	.0361	.1008	.1563	.1755	.1606	.1277	.0916	.0607	.0378
6	.0035	.0120	.0504	.1042	.1462	.1606	.1490	.1221	.0911	.0631
7	.0008	.0034	.0216	.0595	.1044	.1377	.1490	.1396	.1171	.0901
8	.0001	.0009	.0081	.0298	.0653	.1033	.1304	.1396	.1318	.1126
9	.0000	.0002	.0027	.0132	.0363	.0688	.1014	.1241	.1318	.1251
10	.0000	.0000	.0008	.0053	.0181	.0413	.0710	.0993	.1186	.1251
11	.0000	.0000	.0002	.0019	.0082	.0225	.0452	.0722	.0970	.1137
12	.0000	.0000	.0001	.0006	.0034	.0113	.0264	.0481	.0728	.0948
13	.0000	.0000	.0000	.0002	.0013	.0052	.0142	.0296	.0504	.0729
14	.0000	.0000	.0000	.0001	.0005	.0022	.0071	.0169	.0324	.0521
15	.0000	.0000	.0000	.0000	.0002	.0009	.0033	.0090	.0194	.0347
16	.0000	.0000	.0000	.0000	.0000	.0003	.0014	.0045	.0109	.0217
17	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0021	.0058	.0128
18	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0009	.0029	.0071
19	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0014	.0037
20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0006	.0019
21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0009
22	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004
23	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002
24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
25	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Table of the χ^2 Distribution

If X has a χ^2 distribution with m degrees of freedom, this table gives the value of x such that $\Pr(X \le x) = p$, the p quantile of X.

	p								
m	.005	.01	.025	.05	.10	.20	.25	.30	.40
1	.0000	.0002	.0010	.0039	.0158	.0642	.1015	.1484	.2750
2	.0100	.0201	.0506	.1026	.2107	.4463	.5754	.7133	1.022
3	.0717	.1148	.2158	.3518	.5844	1.005	1.213	1.424	1.869
4	.2070	.2971	.4844	.7107	1.064	1.649	1.923	2.195	2.753
5	.4117	.5543	.8312	1.145	1.610	2.343	2.675	3.000	3.655
6	.6757	.8721	1.237	1.635	2.204	3.070	3.455	3.828	4.570
7	.9893	1.239	1.690	2.167	2.833	3.822	4.255	4.671	5.493
8	1.344	1.647	2.180	2.732	3.490	4.594	5.071	5.527	6.423
9	1.735	2.088	2.700	3.325	4.168	5.380	5.899	6.393	7.357
10	2.156	2.558	3.247	3.940	4.865	6.179	6.737	7.267	8.295
11	2.603	3.053	3.816	4.575	5.578	6.989	7.584	8.148	9.237
12	3.074	3.571	4.404	5.226	6.304	7.807	8.438	9.034	10.18
13	3.565	4.107	5.009	5.892	7.042	8.634	9.299	9.926	11.13
14	4.075	4.660	5.629	6.571	7.790	9.467	10.17	10.82	12.08
15	4.601	5.229	6.262	7.261	8.547	10.31	11.04	11.72	13.03
16	5.142	5.812	6.908	7.962	9.312	11.15	11.91	12.62	13.98
17	5.697	6.408	7.564	8.672	10.09	12.00	12.79	13.53	14.94
18	6.265	7.015	8.231	9.390	10.86	12.86	13.68	14.43	15.89
19	6.844	7.633	8.907	10.12	11.65	13.72	14.56	15.35	16.85
20	7.434	8.260	9.591	10.85	12.44	14.58	15.45	16.27	17.81
21	8.034	8.897	10.28	11.59	13.24	15.44	16.34	17.18	18.77
22	8.643	9.542	10.98	12.34	14.04	16.31	17.24	18.10	19.73
23	9.260	10.20	11.69	13.09	14.85	17.19	18.14	19.02	20.69
24	9.886	10.86	12.40	13.85	15.66	18.06	19.04	19.94	21.65
25	10.52	11.52	13.12	14.61	16.47	18.94	19.94	20.87	22.62
30	13.79	14.95	16.79	18.49	20.60	23.36	24.48	25.51	27.44
40	20.71	22.16	24.43	26.51	29.05	32.34	33.66	34.87	36.16
50	27.99	29.71	32.36	34.76	37.69	41.45	42.94	44.31	46.86
60	35.53	37.48	40.48	43.19	46.46	50.64	52.29	53.81	56.62
70	43.27	45.44	48.76	51.74	55.33	59.90	61.70	63.35	66.40
80	51.17	53.54	57.15	60.39	64.28	69.21	71.14	72.92	76.19
90	59.20	61.75	65.65	69.13	73.29	78.56	80.62	82.51	85.99
100	67.33	70.06	74.22	77.93	82.86	87.95	90.13	92.13	95.81

Table of the χ^2 Distribution *(continued)*

	p								
.50	.60	.70	.75	.80	.90	.95	.975	.99	.995
.4549	.7083	1.074	1.323	1.642	2.706	3.841	5.024	6.635	7.879
1.386	1.833	2.408	2.773	3.219	4.605	5.991	7.378	9.210	10.60
2.366	2.946	3.665	4.108	4.642	6.251	7.815	9.348	11.34	12.84
3.357	4.045	4.878	5.385	5.989	7.779	9.488	11.14	13.28	14.86
4.351	5.132	6.064	6.626	7.289	9.236	11.07	12.83	15.09	16.75
5.348	6.211	7.231	7.841	8.558	10.64	12.59	14.45	16.81	18.55
6.346	7.283	8.383	9.037	9.803	12.02	14.07	16.01	18.48	20.28
7.344	8.351	9.524	10.22	11.03	13.36	15.51	17.53	20.09	21.95
8.343	9.414	10.66	11.39	12.24	14.68	16.92	19.02	21.67	23.59
9.342	10.47	11.78	12.55	13.44	15.99	18.31	20.48	23.21	25.19
10.34	11.53	12.90	13.70	14.63	17.27	19.68	21.92	24.72	26.76
11.34	12.58	14.01	14.85	15.81	18.55	21.03	23.34	26.22	28.30
12.34	13.64	15.12	15.98	16.98	19.81	22.36	24.74	27.69	29.82
13.34	14.69	16.22	17.12	18.15	21.06	23.68	26.12	29.14	31.32
14.34	15.73	17.32	18.25	19.31	22.31	25.00	27.49	30.58	32.80
15.34	16.78	18.42	19.37	20.47	23.54	26.30	28.85	32.00	34.27
16.34	17.82	19.51	20.49	21.61	24.77	27.59	30.19	33.41	35.72
17.34	18.87	20.60	21.60	22.76	25.99	28.87	31.53	34.81	37.16
18.34	19.91	21.69	22.72	23.90	27.20	30.14	32.85	36.19	38.58
19.34	20.95	22.77	23.83	25.04	28.41	31.41	34.17	37.57	40.00
20.34	21.99	23.86	24.93	26.17	29.62	32.67	35.48	38.93	41.40
21.34	23.03	24.94	26.04	27.30	30.81	33.92	36.78	40.29	42.80
22.34	24.07	26.02	27.14	28.43	32.01	35.17	38.08	41.64	44.18
23.34	25.11	27.10	28.24	29.55	33.20	36.42	39.36	42.98	45.56
24.34	26.14	28.17	29.34	30.68	34.38	37.65	40.65	44.31	46.93
29.34	31.32	33.53	34.80	36.25	40.26	43.77	46.98	50.89	53.67
39.34	41.62	44.16	45.62	47.27	51.81	55.76	59.34	63.69	66.77
49.33	51.89	54.72	56.33	58.16	63.17	67.51	71.42	76.15	79.49
59.33	62.13	65.23	66.98	68.97	74.40	79.08	83.30	88.38	91.95
69.33	72.36	75.69	77.58	79.71	85.53	90.53	95.02	100.4	104.2
79.33	82.57	86.12	88.13	90.41	96.58	101.9	106.6	112.3	116.3
89.33	92.76	96.52	98.65	101.1	107.6	113.1	118.1	124.1	128.3
99.33	102.9	106.9	109.1	111.7	118.5	124.3	129.6	135.8	140.2

Table of the *t* Distribution

If X has a t distribution with m degrees of freedom, the table gives the value of x such that $\Pr(X \leq x) = p$.

m	$p = .55$	.60	.65	.70	.75	.80	.85	.90	.95	.975	.99	.995
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947
16	.128	.258	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921
17	.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898
18	.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878
19	.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861
20	.127	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750
40	.126	.255	.388	.529	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704
60	.126	.254	.387	.527	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660
120	.126	.254	.386	.526	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617
∞	.126	.253	.385	.524	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576

Table of the Standard Normal Distribution Function

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}u^2\right) du$$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	0.5000	0.60	0.7257	1.20	0.8849	1.80	0.9641	2.40	0.9918
0.01	0.5040	0.61	0.7291	1.21	0.8869	1.81	0.9649	2.41	0.9920
0.02	0.5080	0.62	0.7324	1.22	0.8888	1.82	0.9656	2.42	0.9922
0.03	0.5120	0.63	0.7357	1.23	0.8907	1.83	0.9664	2.43	0.9925
0.04	0.5160	0.64	0.7389	1.24	0.8925	1.84	0.9671	2.44	0.9927
0.05	0.5199	0.65	0.7422	1.25	0.8944	1.85	0.9678	2.45	0.9929
0.06	0.5239	0.66	0.7454	1.26	0.8962	1.86	0.9686	2.46	0.9931
0.07	0.5279	0.67	0.7486	1.27	0.8980	1.87	0.9693	2.47	0.9932
0.08	0.5319	0.68	0.7517	1.28	0.8997	1.88	0.9699	2.48	0.9934
0.09	0.5359	0.69	0.7549	1.29	0.9015	1.89	0.9706	2.49	0.9936
0.10	0.5398	0.70	0.7580	1.30	0.9032	1.90	0.9713	2.50	0.9938
0.11	0.5438	0.71	0.7611	1.31	0.9049	1.91	0.9719	2.52	0.9941
0.12	0.5478	0.72	0.7642	1.32	0.9066	1.92	0.9726	2.54	0.9945
0.13	0.5517	0.73	0.7673	1.33	0.9082	1.93	0.9732	2.56	0.9948
0.14	0.5557	0.74	0.7704	1.34	0.9099	1.94	0.9738	2.58	0.9951
0.15	0.5596	0.75	0.7734	1.35	0.9115	1.95	0.9744	2.60	0.9953
0.16	0.5636	0.76	0.7764	1.36	0.9131	1.96	0.9750	2.62	0.9956
0.17	0.5675	0.77	0.7794	1.37	0.9147	1.97	0.9756	2.64	0.9959
0.18	0.5714	0.78	0.7823	1.38	0.9162	1.98	0.9761	2.66	0.9961
0.19	0.5753	0.79	0.7852	1.39	0.9177	1.99	0.9767	2.68	0.9963
0.20	0.5793	0.80	0.7881	1.40	0.9192	2.00	0.9773	2.70	0.9965
0.21	0.5832	0.81	0.7910	1.41	0.9207	2.01	0.9778	2.72	0.9967
0.22	0.5871	0.82	0.7939	1.42	0.9222	2.02	0.9783	2.74	0.9969
0.23	0.5910	0.83	0.7967	1.43	0.9236	2.03	0.9788	2.76	0.9971
0.24	0.5948	0.84	0.7995	1.44	0.9251	2.04	0.9793	2.78	0.9973
0.25	0.5987	0.85	0.8023	1.45	0.9265	2.05	0.9798	2.80	0.9974
0.26	0.6026	0.86	0.8051	1.46	0.9279	2.06	0.9803	2.82	0.9976
0.27	0.6064	0.87	0.8079	1.47	0.9292	2.07	0.9808	2.84	0.9977
0.28	0.6103	0.88	0.8106	1.48	0.9306	2.08	0.9812	2.86	0.9979
0.29	0.6141	0.89	0.8133	1.49	0.9319	2.09	0.9817	2.88	0.9980
0.30	0.6179	0.90	0.8159	1.50	0.9332	2.10	0.9821	2.90	0.9981
0.31	0.6217	0.91	0.8186	1.51	0.9345	2.11	0.9826	2.92	0.9983
0.32	0.6255	0.92	0.8212	1.52	0.9357	2.12	0.9830	2.94	0.9984
0.33	0.6293	0.93	0.8238	1.53	0.9370	2.13	0.9834	2.96	0.9985
0.34	0.6331	0.94	0.8264	1.54	0.9382	2.14	0.9838	2.98	0.9986
0.35	0.6368	0.95	0.8289	1.55	0.9394	2.15	0.9842	3.00	0.9987
0.36	0.6406	0.96	0.8315	1.56	0.9406	2.16	0.9846	3.05	0.9989
0.37	0.6443	0.97	0.8340	1.57	0.9418	2.17	0.9850	3.10	0.9990
0.38	0.6480	0.98	0.8365	1.58	0.9429	2.18	0.9854	3.15	0.9992
0.39	0.6517	0.99	0.8389	1.59	0.9441	2.19	0.9857	3.20	0.9993
0.40	0.6554	1.00	0.8413	1.60	0.9452	2.20	0.9861	3.25	0.9994
0.41	0.6591	1.01	0.8437	1.61	0.9463	2.21	0.9864	3.30	0.9995
0.42	0.6628	1.02	0.8461	1.62	0.9474	2.22	0.9868	3.35	0.9996
0.43	0.6664	1.03	0.8485	1.63	0.9485	2.23	0.9871	3.40	0.9997
0.44	0.6700	1.04	0.8508	1.64	0.9495	2.24	0.9875	3.45	0.9997
0.45	0.6736	1.05	0.8531	1.65	0.9505	2.25	0.9878	3.50	0.9998
0.46	0.6772	1.06	0.8554	1.66	0.9515	2.26	0.9881	3.55	0.9998
0.47	0.6808	1.07	0.8577	1.67	0.9525	2.27	0.9884	3.60	0.9998
0.48	0.6844	1.08	0.8599	1.68	0.9535	2.28	0.9887	3.65	0.9999
0.49	0.6879	1.09	0.8621	1.69	0.9545	2.29	0.9890	3.70	0.9999
0.50	0.6915	1.10	0.8643	1.70	0.9554	2.30	0.9893	3.75	0.9999
0.51	0.6950	1.11	0.8665	1.71	0.9564	2.31	0.9896	3.80	0.9999
0.52	0.6985	1.12	0.8686	1.72	0.9573	2.32	0.9898	3.85	0.9999
0.53	0.7019	1.13	0.8708	1.73	0.9582	2.33	0.9901	3.90	1.0000
0.54	0.7054	1.14	0.8729	1.74	0.9591	2.34	0.9904	3.95	1.0000
0.55	0.7088	1.15	0.8749	1.75	0.9599	2.35	0.9906	4.00	1.0000
0.56	0.7123	1.16	0.8770	1.76	0.9608	2.36	0.9909		
0.57	0.7157	1.17	0.8790	1.77	0.9616	2.37	0.9911		
0.58	0.7190	1.18	0.8810	1.78	0.9625	2.38	0.9913		
0.59	0.7224	1.19	0.8830	1.79	0.9633	2.39	0.9916		

Table of the 0.95 Quantile of the *F* Distribution

If X has an F distribution with m and n degrees of freedom, the table gives the value of x such that $\Pr(X \le x) = 0.95$.

	m																
n	1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	245.9	248.0	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.70	2.66	2.62	2.58	2.54
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.25	2.20	2.16	2.11	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.04	1.99	1.95	1.90	1.84
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.75	1.66	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.67	1.57	1.46	1.39	1.32	1.22	1.00

Table of the 0.975 Quantile of the *F* Distribution

If X has an F distribution with m and n degrees of freedom, the table gives the value of x such that $\Pr(X \le x) = 0.975$.

	m																
n	1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	984.9	993.1	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43	39.45	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25	14.17	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.56	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.33	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.17	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.47	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	4.00	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.67	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.42	3.31	3.26	3.20	3.14	3.08
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.76	2.64	2.59	2.52	2.46	2.40
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.46	2.35	2.29	2.22	2.16	2.09
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.20	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.18	2.07	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.06	1.94	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	1.94	1.82	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.83	1.71	1.57	1.48	1.39	1.27	1.00

This page intentionally left blank

Answers to Odd-Numbered Exercises

Note: Answers are not provided for exercises that request a proof, a derivation, or a graph.

Chapter 1

Section 1.4

7. (a) $\{x\colon x < 1 \text{ or } x > 5\}$; (b) $\{x\colon 1 \le x \le 7\}$; (c) B; (d) $\{x\colon 0 < x < 1 \text{ or } x > 7\}$; (e) $\emptyset$. **11.** (a) $S = \{(x, y) : 0 \le x \le 5 \text{ and } 0 \le y \le 5\}$. (b) $A = \{(x, y) \in S : x + y \ge 6\}$, $B = \{(x, y) \in S : x = y\}$, $C = \{(x, y) \in S : x > y\}$, $D = \{(x, y) \in S : 5 < x + y < 6\}$. (c) $A^c \cap D^c \cap B$. (d) $A^c \cap B^c \cap C^c$.

Section 1.5

1. $\frac{2}{5}$. **3.** (a) $\frac{1}{2}$; (b) $\frac{1}{6}$; (c) $\frac{3}{8}$. **5.** 0.4. **7.** 0.4 if $A \subset B$ and 0.1 if $\Pr(A \cup B) = 1$. **11.** (a) $1 - \frac{\pi}{4}$; (b) $\frac{3}{4}$; (c) $\frac{2}{3}$; (d) 0.

Section 1.6

1. $\frac{1}{2}$. **3.** $\frac{2}{3}$. **5.** $\frac{4}{7}$. **7.** $\Pr(Aa) = \Pr(aa) = \frac{1}{2}$.

Section 1.7

1. 14. **3.** 5!. **5.** $\frac{5}{18}$. **7.** $\frac{20!}{8!20^{12}}$. **9.** $\frac{(3!)^2}{6!}$.

Section 1.8

1. $\binom{20}{10}$. **3.** They are equal. **5.** This number is $\binom{4251}{97}$, and therefore it must be an integer. **7.** $\frac{n+1-k}{\binom{n}{k}}$. **9.** $\frac{n+1}{\binom{2n}{n}}$. **11.** $\frac{\binom{98}{10}}{\binom{100}{12}}$. **13.** $\frac{\binom{20}{6}+\binom{20}{10}}{\binom{24}{10}}$. **17.** $\frac{4\binom{13}{4}}{\binom{52}{4}}$. **21.** $\binom{365+k}{k}$.

Section 1.9

1. $\binom{21}{7, 7, 7}$. **3.** $\binom{300}{5, 8, 287}$. **5.** $\frac{1}{6^n}\binom{n}{n_1, n_2, \ldots, n_6}$. **7.** $\frac{\binom{12}{6, 2, 4}\binom{13}{4, 6, 3}}{\binom{25}{10, 8, 7}}$. **9.** $\frac{4!}{\binom{52}{13, 13, 13, 13}}$.

Section 1.10

1. $3\frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}} - 3\frac{\binom{4}{2}\cdot\binom{48}{3, 3, 42}}{\binom{52}{5, 5, 42}}$. **3.** 45 percent. **5.** $\frac{3}{8}$. **7.** $1 - \frac{1}{\binom{100}{15}}\left\{\left[\binom{90}{15} + \binom{80}{15} + \binom{70}{15} + \binom{60}{15}\right] - \left[\binom{70}{15} + \binom{60}{15} + \binom{50}{15} + \binom{50}{15} + \binom{40}{15} + \binom{30}{15}\right] + \left[\binom{40}{15} + \binom{30}{15} + \binom{20}{15}\right]\right\}$. **9.** $n = 10$. **11.** $\frac{\binom{5}{r}\binom{5}{5-r}}{\binom{10}{5}}$, where $r = \frac{x}{2}$ and $x = 0, 2, \ldots, 10$.

Section 1.12

1. No. **3.** $\frac{\binom{250}{18}\cdot\binom{100}{12}}{\binom{350}{30}}$. **5.** 0.3120 **7.** $\frac{1}{\binom{r+w}{r}}$. **9.** $\frac{\binom{7}{j}\binom{3}{5-j}}{\binom{10}{5}}$, where $k = 2j - 2$ and $j = 2, 3, 4, 5$. **13.** (d) $\frac{\binom{n-k+1}{k}}{\binom{n}{k}}$.

Chapter 2

Section 2.1

1. $\Pr(A)/\Pr(B)$. **3.** $\Pr(A)$. **5.** $\frac{r(r+k)(r+2k)b}{(r+b)(r+b+k)(r+b+2k)(r+b+3k)}$. **7.** $\frac{1}{3}$. **9.** (a) $\frac{3}{4}$; (b) $\frac{3}{5}$. **13.** 0.44. **15.** 0.47.

Section 2.2

1. $\Pr(A^c)$. **5.** $1 - \frac{1}{10^6}$. **7.** (a) 0.92; (b) 0.8696 **9.** $\frac{1}{7}$. **11.** (a) 0.2617. **13.** $10(0.01)(0.99)^9$. **15.** $n > \frac{\log(0.2)}{\log(0.99)}$. **17.** $\frac{1}{12}$. **19.** $[(0.8)^{10} + (0.7)^{10}] - [(0.2)^{10} + (0.3)^{10}]$. **23.** (a) 0.2215; (b) 0.0234.

Section 2.3

3. 0.301. **5.** $\frac{18}{59}$. **7.** (a) $0, \frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10}$; (b) $\frac{3}{4}$; (c) $\frac{1}{4}$. **11.** 1/4. **13.** (a) 1/9; (b) 1. **15.** 0.274.

Section 2.4

3. Condition (a). **5.** $i \ge 198$. **9.** $\frac{2}{3}$.

Section 2.5

3. $\frac{11}{12}$. **5.** $\frac{1}{\binom{10}{3}}$. **7.** Always. **9.** $\frac{1}{6}$. **11.** $1-\left(\frac{49}{50}\right)^{50}$. **13.** (a) 0.93; (b) 0.38. **15.** $\frac{4}{81}$. **17.** 0.067.
19. $p_1+p_2+p_3-p_1p_2-p_2p_3-p_1p_3+p_1p_2p_3$, where

$$p_1=\frac{\binom{6}{1}}{\binom{8}{3}},\quad p_2=\frac{\binom{6}{2}}{\binom{8}{4}},\quad p_3=\frac{\binom{6}{3}}{\binom{8}{5}}.$$

21. $\Pr(A \text{ wins})=\frac{4}{7}$; $\Pr(B \text{ wins})=\frac{2}{7}$; $\Pr(C \text{ wins})=\frac{1}{7}$. **23.** 0.372. **25.** (a) 0.659; (b) 0.051. **27.** $\frac{1-\left(\frac{1}{2}\right)^{n-1}}{1-\left(\frac{1}{2}\right)^{n}}$.
29. (a) $\frac{1-p_0-p_1}{1-p_0}$, where $p_0=\frac{\binom{48}{13}}{\binom{52}{13}}$ and $p_1=\frac{4\binom{48}{12}}{\binom{52}{13}}$. (b) $1-p_1$. Also, $\frac{\binom{3}{1}\binom{48}{11}+\binom{3}{2}\binom{48}{10}+\binom{48}{9}}{\binom{51}{12}}=0.5612$ **33.** $\frac{7}{9}$. **35.** (a) The second condition; (b) The first condition; (c) Equal probability under both conditions.

Chapter 3

Section 3.1

1. $\frac{6}{11}$. **3.** $f(0)=\frac{1}{6}$, $f(1)=\frac{5}{18}$, $f(2)=\frac{2}{9}$, $f(3)=\frac{1}{6}$, $f(4)=\frac{1}{9}$, $f(5)=\frac{1}{18}$. **5.** $f(x)=\begin{cases}\frac{\binom{7}{x}\binom{3}{5-x}}{\binom{10}{5}} & \text{for } x=2,3,4,5,\\ 0 & \text{otherwise}\end{cases}$
7. 0.806. **9.** 1/2.

Section 3.2

1. 4/9. **3.** (a) $\frac{1}{2}$; (b) $\frac{13}{27}$; (c) $\frac{2}{27}$. **5.** (a) $t=2$; (b) $t=\sqrt{8}$. **7.** $f(x)=\begin{cases}\frac{1}{10} & \text{for } -2\le x\le 8,\\ 0 & \text{otherwise,}\end{cases}$ and probability is $\frac{7}{10}$. **13.** 0.0045.

Section 3.3

5. $f(x)=(2/9)x$ for $0\le x\le 3$; $f(x)=0$ otherwise. **7.** $F(x)=\begin{cases}0 & \text{for } x<-2,\\ \frac{1}{10}(x+2) & \text{for } -2\le x\le 8,\\ 1 & \text{for } x>9.\end{cases}$ **11.** $F^{-1}(p)=3p^{1/2}$.
13. 10.2. **15.** $F(x)=x^2$ for $0<x<1$.

Section 3.4

1. (a) 0.5; (b) 0.75. **3.** (a) $\frac{1}{40}$; (b) $\frac{1}{20}$; (c) $\frac{7}{40}$; (d) $\frac{7}{10}$. 5.(a) $\frac{5}{4}$; (b) $\frac{79}{256}$; (c) $\frac{13}{16}$; (d) 0. **7.** (a) 0.55; (b) 0.8.
9. 0.63505. **11.** (a) 0.273; (b) 0.513.

Section 3.5

1. Uniform on the interval $[a, b]$ and uniform on the interval $[c, d]$. **3.** (a) $f_1(x)=\begin{cases}\frac{1}{2} & \text{for } 0\le x\le 2,\\ 0 & \text{otherwise}\end{cases}$ $f_2(y)=\begin{cases}3y^2 & \text{for } 0\le y\le 1,\\ 0 & \text{otherwise}\end{cases}$ (b) Yes; (c) Yes. **5.** (a) $f(x,y)=\begin{cases}p_x p_y & \text{for } x=0,1,2,3 \text{ and } y=0,1,2,3,\\ 0 & \text{otherwise}\end{cases}$ (b) 0.3; (c) 0.35.
7. Yes. **9.** (a) $f(x,y)=\begin{cases}\frac{1}{6} & \text{for } (x,y)\in S,\\ 0 & \text{otherwise}\end{cases}$
$f_1(x)=\begin{cases}\frac{1}{2} & \text{for } 0\le x\le 2,\\ 0 & \text{otherwise}\end{cases}$ $f_2(y)=\begin{cases}\frac{1}{3} & \text{for } 1\le y\le 4,\\ 0 & \text{otherwise}\end{cases}$ (b) Yes. **11.** $\frac{11}{36}$. **15.** (b) $f_1(x)=1/3$ for $1<x<3$, $f_1(x)=1/6$ for $6<x<8$, and $f_1(x)=0$ otherwise; $f_2(y)=1$ for $0<y<1$ and $f_2(y)=0$ otherwise.

Section 3.6

1. For $-1<y<1$, $g_1(x|y)=\begin{cases}1.5x^2(1-y^2)^{-3/2} & \text{for } -(1-y^2)^{1/2}<x<(1-y^2)^{1/2},\\ 0 & \text{otherwise.}\end{cases}$

3. (a) For $-2 < x < 4$, $g_2(y|x) = \begin{cases} \frac{1}{2[9-(x-1)^2]^{1/2}} & \text{for } (y+2)^2 < 9-(x-1)^2, \\ 0 & \text{otherwise.} \end{cases}$
(b) $\frac{2-\sqrt{2}}{4}$. **5.** (a) For $0 < y < 1$, $g_1(x|y) = \begin{cases} \frac{-1}{(1-x)\log(1-y)} & \text{for } 0 < x < y, \\ 0 & \text{otherwise} \end{cases}$ (b) $\frac{1}{2}$. **7.** (a) For $0 < x < 2$, $g_2(y|x) = \begin{cases} \frac{4-2x-y}{2(2-x)^2} & \text{for } 0 < y < 4-2x, \\ 0 & \text{otherwise} \end{cases}$ (b) $\frac{1}{9}$. **9.** (a) $f_1(x) = \begin{cases} \frac{1}{2}x(2+3x) & \text{for } 0 < x < 1, \\ 0 & \text{otherwise} \end{cases}$ (b) $\frac{8}{11}$. **13.** $g_1(1|1) = 0.5506$, $g_1(1|2) = 0.6561$, $g_1(1|3) = 0.4229$, $g_1(1|4) = 0.2952$. $g_1(0|y) = 1 - g_1(1|y)$ for $y = 1, 2, 3, 4$.

Section 3.7

1. (a) 1/3; (b) $(x_1 + 3x_3 + 1)/3$ for $0 \le x_i \le 1$ $(i = 1, 3)$; (c) 5/13. **3.** (a) 6;
(b) $f_{13}(x_1, x_3) = \begin{cases} 3e^{-(x_1+3x_3)} & \text{for } x_i > 0 (i = 1, 3), \\ 0 & \text{otherwise} \end{cases}$ (c) $1 - \frac{1}{e}$.
5. (a) $\prod_{i=1}^{n} p_i$; (b) $1 - \prod_{i=1}^{n}(1 - p_i)$. **7.** $\sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$, where $p = \int_a^b f(x)\,dx$.

Section 3.8

1. $g(y) = \begin{cases} 3(1-y)^{1/2}/2 & \text{for } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$ **3.** $G(y) = 1 - (1-y)^{1/2}$ for $0 < y < 1$; $g(y) = \begin{cases} \frac{1}{2(1-y)^{1/2}} & \text{for } 0 < y < 1, \\ 0 & \text{otherwise} \end{cases}$
7. (a) $g(y) = \begin{cases} \frac{1}{2}y^{-1/2} & \text{for } 0 < y < 1, \\ 0 & \text{otherwise} \end{cases}$ (b) $g(y) = \begin{cases} \frac{1}{3}|y|^{-2/3} & \text{for } -1 < y < 0, \\ 0 & \text{otherwise} \end{cases}$ (c) $g(y) = \begin{cases} 2y & \text{for } 0 < y < 1, \\ 0 & \text{otherwise} \end{cases}$
9. $Y = 2X^{1/3}$. **13.** $f(t) = \begin{cases} 2e^{-2/t}/t^2 & \text{for } t > 0, \\ 0 & \text{otherwise.} \end{cases}$ **17.** (a) $r(x) = 0$ for $x \le 100$, $r(x) = x - 100$ for $100 < x \le 5100$, $r(x) = 5000$ for $x > 5100$; (b) $G(y) = 0$ for $y < 0$, $G(y) = 1 - 1/(y + 101)$ for $0 \le y < 5000$, $G(y) = 1$ for $y \ge 5000$.

Section 3.9

1. $g(y) = \begin{cases} y & \text{for } 0 < y \le 1, \\ 2-y & \text{for } 1 < y < 2, \\ 0 & \text{otherwise} \end{cases}$ **3.** $g(y_1, y_2, y_3) = \begin{cases} 8y_3(y_1y_2)^{-1} & \text{for } 0 < y_3 < y_2 < y_1 < 1, \\ 0 & \text{otherwise.} \end{cases}$
5. $g(z) = \begin{cases} \frac{1}{3}(z+1) & \text{for } 0 < z \le 1, \\ \frac{1}{3z^3}(z+1) & \text{for } z > 1, \\ 0 & \text{for } z \le 0. \end{cases}$ **7.** $g(y) = \frac{1}{2}e^{-|y|}$ for $-\infty < y < \infty$. **9.** $(0.8)^n - (0.7)^n$. **11.** $\left(\frac{1}{3}\right)^n + \left(\frac{2}{3}\right)^n$.
13. $f(z) = \begin{cases} \frac{n(n-1)}{8}\left(\frac{z}{8}\right)^{n-2}\left(1 - \frac{z}{8}\right) & \text{for } -3 < z < 5, \\ 0 & \text{otherwise} \end{cases}$ **19.** ye^{-y} for $y > 0$.

Section 3.10

1. (a) (1/2,1/2); (b) $\begin{pmatrix} 5/9 & 4/9 \\ 4/9 & 5/9 \end{pmatrix}$ **3.** (a) 0.667; (b) 0.666. **5.** (a) 0.38; (b) 0.338; (c) 0.3338. **7.** (a) 0.632; (b) 0.605.
9. (a) $\frac{1}{8}$; (b) $\frac{1}{8}$. **11.** (a) $\frac{40}{81}$; (b) $\frac{41}{81}$.
13.

	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
HHH	0	1	0	0	0	0	0	0
HHT	0	0	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$	0
HTH	0	0	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0
THH	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0	0
TTH	0	0	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0
THT	0	0	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$	0
HTT	0	0	0	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$
TTT	0	0	0	0	1	0	0	0

17. (a) $\{Aa, Aa\}$ has probability 1; (b) $\{Aa, Aa\}$, $\{Aa, aa\}$, and $\{aa, aa\}$ have, respectively, probabilities 0.04, 0.32, and 0.64. **19.** (2/3, 1/3).

Section 3.11

3. $f(x) = \begin{cases} \frac{2}{5} & \text{for } 0 < x < 1, \\ \frac{3}{5} & \text{for } 1 < x < 2, \\ 0 & \text{otherwise} \end{cases}$ **5.** $\frac{\pi}{4}$. **7.** $1 - \frac{1}{2^{p-1}} + \frac{1}{2^{2p-1}}$. **9.** $\frac{1}{10}$. **11.** $Y = 5(1 - e^{-2X})$ or $Y = 5e^{-2X}$. **13.** The sets (c) and (d). **15.** 0.3715. **17.** $f_2(y) = -9y^2 \log y$ for $0 < y < 1$. $g_1(x|y) = -\frac{1}{x \log y}$ for $0 < y < x < 1$. **19.** $f_1(x) = 3(1-x)^2$ for $0 < x < 1$, $f_2(y) = 6y(1-y)$ for $0 < y < 1$, $f_3(z) = 3z^2$ for $0 < z < 1$.
21. (a) $g(u, v) = \begin{cases} ve^{-v} & \text{for } 0 < u < 1, v > 0, \\ 0 & \text{otherwise} \end{cases}$ (b) Yes. **23.** $h(y_1|y_n) = \frac{(n-1)(e^{-y_1} - e^{-y_n})^{n-2} e^{-y_1}}{(1-e^{-y_n})^{n-1}}$ for $0 < y_1 < y_n$.
25. (a) $2\epsilon f_2(y)$; (b) $2\epsilon \int_{-\infty}^{x} f(s, y)ds$.
27.

		Players in game $n+1$		
		(A, B)	(A, C)	(B, C)
Players in	(A, B)	0	0.3	0.7
game n	(A, C)	0.6	0	0.4
	(B, C)	0.8	0.2	0

29. (0.4220, 0.2018, 0.3761).

Chapter 4

Section 4.1

1. $(a+b)/2$. **3.** 18.92. **5.** 4.867. **9.** $\frac{3}{4}$. **11.** $\frac{1}{n+1}$ and $\frac{n}{n+1}$. **13.** \$11.61. **15.** \$25.

Section 4.2

1. \$5. **3.** $\frac{1}{2}$. **5.** $n \int_a^b f(x)\,dx$. **7.** $c\left(\frac{5}{4}\right)^n$. **9.** $n(2p-1)$. **11.** $2k$.

Section 4.3

1. 1/12. **3.** $\frac{1}{12}(b-a)^2$. **7.** (a) 6; (b) 39. **9.** $(n^2-1)/12$. **11.** 0.5. **13.** 1.

Section 4.4

1. 0. **3.** 1. **7.** $\mu = \frac{1}{2}, \sigma^2 = \frac{3}{4}$. **9.** $E(Y) = c\mu$; $\text{Var}(Y) = c(\sigma^2 + \mu^2)$. **11.** $f(1) = \frac{1}{5}$; $f(4) = \frac{2}{5}$; $f(8) = \frac{2}{5}$. **17.** 2.

Section 4.5

3. $m = \log 2$. **5.** (a) $\frac{1}{2}(\mu_f + \mu_g)$; (b) Any number m such that $1 \le m \le 2$. **7.** (a) $\frac{7}{12}$; (b) $\frac{1}{2}(\sqrt{5} - 1)$.
9. (a) 0.1; (b) 1. **11.** Y.

Section 4.6

1. 0. **11.** The value of $\rho(X, Y)$ would be less than -1. **13.** (a) 11; (b) 51. **15.** $n + \frac{n(n-1)}{4}$.

Section 4.7

1. 0.00576, 7% of the marginal M.S.E. **5.** $1 - \frac{1}{2^n}$. **7.** $E(Y|X) = \frac{3X+2}{3(2X+1)}$; $\text{Var}(Y|X) = \frac{1}{36}\left[3 - \frac{1}{(2X+1)^2}\right]$.
9. $\frac{1}{12} - \frac{\log 3}{144}$. **13.** (a) $\frac{3}{5}$; (b) $\frac{\sqrt{29}-3}{4}$. **15.** (a) $\frac{18}{31}$; (b) $\frac{\sqrt{5}-1}{2}$.

Section 4.8

1. $\alpha > 1.111$. **3.** Z. **5.** $\frac{2}{3}$ **7.** p. **9.** $a = 1$ if $p > \frac{1}{2}$; $a = 0$ if $p < \frac{1}{2}$; a can be chosen arbitrarily if $p = \frac{1}{2}$.
11. $b = 0$ if $p \le \frac{1}{2}$; $b = (2p-1)A$ if $p > \frac{1}{2}$. **13.** $b = A$ if $p > \frac{1}{2}$; $b = 0$ if $p < \frac{1}{2}$; b can be chosen arbitrarily if $p = \frac{1}{2}$.
15. $x_0 > \frac{4}{(\alpha+1)^{1/\alpha}}$. **17.** Continue to promote.

Section 4.9

5. $a = \pm\frac{1}{\sigma}$, $b = -a\mu$. **7.** $\frac{3}{2}$. **11.** Order an amount s such that $\int_0^s f(x)\,dx = \frac{g}{g+c}$. **13.** (a) and (b) $E(Z) = 29$; $\text{Var}(Z) = 109$. (c) $E(Z) = 29$; $\text{Var}(Z) = 94$. **17.** 1. **21.** $-\frac{1}{2}$. **25.** (a) 0.1333. (b) 0.1414. **29.** $a = pm$.

Chapter 5

Section 5.2

1. Bernoulli with parameter $\frac{1}{3}$. **3.** 0.377. **5.** 0.5000. **7.** $\frac{113}{64}$. **9.** $\frac{k}{n}$.
11. $n(n-1)p^2$. **13.** 0.4957 **15.** 1110, 4.64×10^{-171}.

Section 5.3

1. 8.39×10^{-8}. **3.** $E(\overline{X}) = \frac{1}{3}$; $\text{Var}(\overline{X}) = \frac{8}{441}$. **5.** $\frac{T-1}{2}$ or $\frac{T+1}{2}$ if T is odd, and $\frac{T}{2}$ if T is even.
7. (a) $\frac{\binom{0.7T}{10}+0.3T\binom{0.7T}{9}}{\binom{T}{10}}$; (b) $(0.7)^{10} + 10(0.3)(0.7)^9$. **9.** 3/128.

Section 5.4

1. 0.5940. **3.** 0.0166. **5.** $\sum_{x=m}^{n}\binom{n}{x}\left(\sum_{i=k+1}^{\infty}\frac{e^{-\lambda}\lambda^i}{i!}\right)^x\left(\sum_{i=0}^{k}\frac{e^{-\lambda}\lambda^i}{i!}\right)^{n-x}$. **7.** $\sum_{x=21}^{\infty}\frac{e^{-30}30^x}{x!}$. **9.** Poisson distribution with mean $p\lambda$. **11.** If λ is not an integer, mode is the greatest integer less than λ. If λ is an integer, modes are λ and $\lambda - 1$. **13.** 0.3476. **15.** $9\lambda e^{-3\lambda}$, for $\lambda > 0$.

Section 5.5

1. (a) 0.0001; (b) 0.**01**. **3.** (a) 150; (b) 4350. **9.** Geometric distribution with parameter $p = 1 - \prod_{i=1}^{n} q_i$.

Section 5.6

1. 0.0, −0.6745, 0.6745, −1.282, 1.282. **3.** Normal with $\mu = 20$ and $\sigma = \frac{20}{9}$. **5.** 0.996. **7.** $(0.1360)^3$. **9.** 0.6827. **11.** $n = 1083$. **13.** 0.3812. **15.** (a) $\frac{\exp\left\{-\frac{1}{2}(x-25)^2\right\}}{\exp\left\{-\frac{1}{2}(x-25)^2\right\}+9\exp\left\{-\frac{1}{2}(x-20)^2\right\}}$; (b) $x > 22.5 + \frac{1}{5}\log 9$. **17.** $f(x) = \frac{1}{(2\pi)^{1/2}\sigma x}\exp\left\{-\frac{1}{2\sigma^2}(\log x - \mu)^2\right\}$ for $x > 0$, and $f(x) = 0$ for $x \le 0$. **19.** $f(\mu) = \frac{1.0013}{(2\pi)^{1/2}}\exp\left\{-\frac{1}{2}(\mu - 8)^2\right\}$ for $5 < \mu < 15$.

21. The lognormal distribution with parameters 4.6 and 10.5. **23.** The lognormal distribution with parameters 3.149 and 2.

Section 5.7

7. $1 - [1 - \exp(-\beta t)]^3$. **9.** $\frac{1}{e}$. **11.** $\left(\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2}\right)\frac{1}{\beta}$. **13.** $1 - e^{-5/2}$.

15. $e^{-5/4}$ **17.** $1 \cdot 3 \cdot 5 \cdots (2n-1)\sigma^{2n}$.

Section 5.8

1. $F^{-1}(p) = p^{1/\alpha}$. **5.** $\frac{\alpha(\alpha+1)\cdots(\alpha+r-1)\beta(\beta+1)\cdots(\beta+s-1)}{(\alpha+\beta)(\alpha+\beta+1)\cdots(\alpha+\beta+r+s-1)}$. **9.** $\alpha = 1/17$, $\beta = 19/17$.

Section 5.9

3. $\frac{2424}{6^5}$. **5.** 0.0501.

Section 5.10

1. 70.57. **3.** 0.1562. **5.** 90 and 36. **7.** $\mu_1 = 4$, $\mu_2 = -2$, $\sigma_1 = 1$, $\sigma_2 = 2$, $\rho = -0.3$. **13.** $\rho = -0.5c/(ab)^{1/2}$, $\sigma_1^2 = 2b/d$, $\sigma_2^2 = 2a/d$, $\mu_1 = (cg - 2be)/d$, $\mu_2 = (ce - 2ag)/d$, where $d = 4ab - c^2$.

Section 5.11

1. $f(x) = 1/(n+1)$ for $x = 0, \ldots, n$. **3.** 0.0404. **7.** $3\mu\sigma^2 + \mu^3$. **9.** 0.8152. **11.** $\frac{15}{7}$. **13.** 0.2202. **15.** (a) Exponential, parameter $\beta = 5$; (b) Gamma, parameters $\alpha = k$ and $\beta = 5$; (c) $e^{-5(k-1)/3}$. **23.** (a) $\rho(X_i, X_j) = -\left(\frac{p_i}{1-p_i} \cdot \frac{p_j}{1-p_j}\right)^{1/2}$, where p_i is the proportion of students in class i; (b) $i = 1$, $j = 2$; (c) $i = 3$, $j = 4$. **25.** Normal with $\mu = -3$ and $\sigma^2 = 16$; $\rho(X, Y) = \frac{1}{2}$.

Chapter 6

Section 6.1

1. $4x$ if $0 < x \le 1/2$, $4 - 4x$ if $1/2 < x < 1$, and 0 otherwise; 0.36; 0.2; look at where each p.d.f. is higher than the other. **3.** 0.9964. The probability looks like it might be increasing to 1.

Section 6.2

5. 25. **13.** (a) Yes; (b) No. **17.** (b) $np(1-p)$ and $knp(1-p/k)$. **21.** (a) $[u \exp(1-u)]^n$; (b) Useless bound.

Section 6.3

1. 0.001 **3.** 0.9938 **5.** $n \ge 542$. **7.** 0.7385. **9.** (a) 0.36; (b) 0.7887. **11.** 0.9938. **13.** Normal with mean θ^3 and variance $\frac{9\theta^4\sigma^2}{n}$. **15.** (c) $n(Y_n^2 - \theta^2)/[2\theta]$ has approximately c.d.f. F^*.

Section 6.4

1. 0.8169. **3.** 0.0012. **5.** 0.9938. **7.** 0.7539.

Section 6.5

1. 8.00. **3.** Without continuity correction, 0.473; with continuity correction, 0.571; exact probability, 0.571. **5.** $\arcsin(\sqrt{\overline{X}_n})$. **9.** 0.1587. **11.** (b) Normal with mean $n/3$ and variance $n/9$.

Chapter 7

Section 7.1

1. $X_1, X_2, \ldots, P$; the X_i are i.i.d. Bernoulli with parameter p given $P = p$. **3.** $Z_1, Z_2, \ldots$ times of hits, parameter β, $Y_k = Z_k - Z_{k-1}$ for $k \ge 2$. **5.** $(\overline{X}_n - 0.98, \overline{X}_n + 0.98)$ has probability 0.95 of containing μ. **7.** Y Poisson with mean λt, parameters λ and p, $X_1, \ldots, X_y$ i.i.d. Bernoulli with parameter p given $Y = y$, $X = X_1 + \cdots + X_y$ (observable).

Section 7.2

1. 0.4516. **3.** $\xi(1.0|X=3) = 0.2456$; $\xi(1.5|X=3) = 0.7544$. **5.** The p.d.f. of the beta distribution with parameters $\alpha = 3$ and $\beta = 6$. **7.** Beta distribution with parameters $\alpha = 4$ and $\beta = 7$. **9.** Beta distribution with parameters $\alpha = 4$ and $\beta = 6$. **11.** Uniform distribution on the interval [11.2, 11.4].

Section 7.3

1. 120. **3.** Beta distribution with parameters $\alpha = 5$ and $\beta = 297$. **5.** Gamma distribution with parameters $\alpha = 16$ and $\beta = 6$. **7.** Normal distribution with mean 69.07 and variance 0.286. **9.** Normal distribution with mean 0 and variance $\frac{1}{5}$. **13.** $n \ge 100$. **17.** $\xi(\theta|x) = \begin{cases} \frac{6(8^6)}{\theta^7} & \text{for } \theta > 8, \\ 0 & \text{for } \theta \le 8. \end{cases}$ **19.** $\frac{\alpha+n}{\beta - \sum_{i=1}^n \log x_i}$ and $\frac{\alpha+n}{\left(\beta - \sum_{i=1}^n \log x_i\right)^2}$. **21.** Gamma distribution with parameters n and $n\overline{x}_n$.

Section 7.4

1. 2/3 and $2^{-1/2}$. **3.** (a) 12 or 13; (b) **0.** **5.** $\frac{8}{3}$. **9.** $n \ge 396$. **13.** $\frac{\alpha+n}{\alpha+n-1}\max(x_0, X_1, \ldots, X_n)$.

Section 7.5

3. $\frac{2}{3}$. **5.** (a) $\hat{\theta} = \overline{x}_n$. **7.** $\hat{\beta} = \frac{1}{\overline{X}_n}$. **9.** $\hat{\theta} = -\frac{n}{\sum_{i=1}^n \log X_i}$. **11.** $\hat{\theta}_1 = \min(X_1, \ldots, X_n)$; $\hat{\theta}_2 = \max(X_1, \ldots, X_n)$.
13. $\hat{\mu}_1 = \overline{X}_n$; $\hat{\mu}_2 = \overline{Y}_n$.

Section 7.6

1. $(\prod_{i=1}^n X_i)^{1/n}$. **3.** $\hat{m} = \overline{X}_n \log 2$. **5.** $\hat{\mu} = \frac{1}{2}[\min\{X_1, \ldots, X_n\} + \max\{X_1, \ldots, X_n\}]$. **7.** $\hat{v} = \Phi\left(\frac{\hat{\mu}-2}{\hat{\sigma}}\right)$.
9. $\overline{X}_n$. **15.** $\hat{\mu} = 6.75$. **17.** $\hat{p} = \frac{2}{5}$. **23.** (a) $\hat{\alpha} = [\overline{x}_n(\overline{x}_n - \overline{x^2}_n)]/[\overline{x^2}_n - \overline{x}_n^2]$, $\hat{\beta} = [(1 - \overline{x}_n)(\overline{x}_n - \overline{x^2}_n)]/[\overline{x^2}_n - \overline{x}_n^2]$.
25. $\hat{\mu}_1 = \overline{X}_n$, $\widehat{\sigma_1^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$, $\hat{\mu}_2 = \hat{\alpha} + \hat{\beta}\hat{\mu}_1$, $\widehat{\sigma_2^2} = \widehat{\sigma_{2.1}^2} + \hat{\beta}^2\widehat{\sigma_1^2}$, $\hat{\rho} = \hat{\beta}\widehat{\sigma_1}/\widehat{\sigma_2}$, where $\hat{\beta} = \sum_{i=1}^{n-k}(Y_i - \overline{Y}_{n-k})(X_i - \overline{X}_{n-k})/\sum_{i=1}^{n-k}(X_i - \overline{X}_{n-k})^2$, $\hat{\alpha} = \overline{Y}_{n-k} - \hat{\beta}\hat{\mu}_1$, and $\widehat{\sigma_{2.1}^2} = \frac{1}{n-k}\sum_{i=1}^{n-k}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$.

Section 7.8

9. Yes. **11.** No. **13.** Yes. **15.** Yes. **17.** Yes.

Section 7.9

3. $R(\theta, \delta_1) = \frac{\theta^2}{3n}$. **5.** $c^* = \frac{n+2}{n+1}$. **7.** (a) $R(\beta, \delta) = (\beta - 3)^2$. **11.** $\hat{\theta} = \delta_0$. **13.** $\left(\frac{n-1}{n}\right)^T$. **15.** $\exp(\overline{X}_n + 0.125)$, $c = 0.125(1 - 3/n)$.

Section 7.10

1. (a) Beta distribution with parameters 11 and 16; (b) 11/27. **3.** $\frac{6}{17}$. **5.** $\frac{\sigma_2^2 b_1 x_1 + \sigma_1^2 b_2 x_2}{\sigma_2^2 b_1^2 + \sigma_1^2 b_2^2}$.
7. (a) $\frac{1}{3}\left(X_1 + \frac{1}{2}X_2 + \frac{1}{3}X_3\right)$; (b) Gamma distribution, parameters $\alpha + 3$ and $\beta + x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3$. **9.** (a) $x + 1$. (b) $x + \log 2$. **11.** $\hat{p} = 2(\hat{\theta} - \frac{1}{4})$, where

$$\hat{\theta} = \begin{cases} \frac{X}{n} & \text{if } \frac{1}{4} \le \frac{X}{n} \le \frac{3}{4}, \\ \frac{1}{4} & \text{if } \frac{X}{n} < \frac{1}{4}, \\ \frac{3}{4} & \text{if } \frac{X}{n} > \frac{3}{4}. \end{cases}$$

13. $2^{1/5}$. **15.** $\min(X_1, \ldots, X_n)$. **17.** $\hat{x}_0 = \min(X_1, \ldots, X_n)$, and $\hat{\alpha} = \left(\frac{1}{n}\sum_{i=1}^n \log x_i - \log \hat{x}_0\right)^{-1}$. **19.** The smallest integer greater than $\frac{x}{p} - 1$. If $\frac{x}{p} - 1$ is itself an integer, both $\frac{x}{p} - 1$ and $\frac{x}{p}$ are M.L.E.'s. **21.** 16.

Chapter 8

Section 8.1

1. $n \ge 29$. **3.** $n \ge 255$. **5.** $n = 10$. **7.** $n \ge 16$. **9.** $1 - G(n/t)$, where $G(\cdot)$ is the c.d.f. of the gamma distribution with parameters n and θ.

Section 8.2

1. 0.1278. **5.** 0.20. **9.** χ^2 distribution with one degree of freedom. **11.** $\frac{2^{1/2}\Gamma[(m+1)/2]}{\Gamma(m/2)}$.

Section 8.3

7. (a) $n = 21$; (b) $n = 13$. **9.** The same for both samples.

Section 8.4

3. $c = \sqrt{3/2}$. **5.** 0.70.

Section 8.5

3. (a) $6.16\sigma^2$; (b) $2.05\sigma^2$; (c) $0.56\sigma^2$; (d) $1.80\sigma^2$; (e) $2.80\sigma^2$; (f) $6.12\sigma^2$.
7. (148.1, 165.6). **9.** (a) (4.7, 5.3); (b) (4.8, 5.2); (d) 0.6; (e) 0.5.
11. Endpoints are $\sin^2\left(\arcsin\sqrt{\overline{x}_n} \pm n^{-1/2}\Phi^{-1}([1+\gamma]/2)\right)$, unless one of the numbers $\arcsin\sqrt{\overline{x}_n} \pm n^{-1/2}\Phi^{-1}([1 +$

$\gamma]/2)$ lies outside of the interval $[0, \pi/2]$.

Section 8.6

5. $\mu_0 = -5$; $\lambda_0 = 4$; $\alpha_0 = 2$; $\beta_0 = 4$. **7.** The conditions imply that $\alpha_0 = \frac{1}{4}$, and $E(\mu)$ exists only for $\alpha_0 > \frac{1}{2}$. **9.** (a) (157.83, 210.07); (b) (152.55, 211.79). **11.** (0.446, 1.530). **13.** (0.724, 3.336). **15.** (a) $\alpha_1 = 7.5$, $\beta_1 = 22.73$, $\lambda_1 = 13$, $\mu_1 = 6.631$; (b) (5.602, 7.660). **17.** $\alpha_1 = 4.5$, $\beta_1 = 0.4831$, $\lambda_1 = 10$, $\mu_1 = 1.379$. **19.** (a) Normal-gamma with hyperparameters $\alpha_1 = 11$, $\beta_1 = 4885.7$, $\lambda_1 = 20.5$, $\mu_1 = 156.7$; (b) (148.7, 164.7).

Section 8.7

1. (a) $g(\theta) = \theta$; (b) $\overline{X}_n$. **3.** $\frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. **5.** $\delta(X) = 2^X$. **11.** (a) All values; (b) $\alpha = \frac{m}{m+4n}$. **15.** (c) $c_0 = \frac{1}{3}(1+\theta_0)$.

Section 8.8

3. $I(\theta) = \frac{1}{\theta}$. **5.** $I(\sigma^2) = \frac{1}{2\sigma^4}$. **9.** $\sqrt{\pi/2}|X|$, $(\pi/2 - 1)\sigma^2$.

Section 8.9

7. (a) For $\alpha(m-1) + 2\beta(n-1) = 1$. (b) $\alpha = \frac{1}{m+n-2}$, $\beta = \frac{1}{2(m+n-2)}$. **9.** $\frac{Y}{2\left[\frac{S_n^2}{n-1}\right]^{1/2}}$. **11.** $\overline{X}_n - c\left[\frac{S_n^2}{n(n-1)}\right]^{1/2}$, where c is the 0.99 quantile of the t distribution with $n-1$ degrees of freedom. **13.** (a) $(\mu_1 - 1.96v_1, \mu_1 + 1.96v_1)$, where μ_1 and v_1 are given by Eqs. (7.3.1) and (7.3.2). **15.** Normal with mean θ and variance θ^2/n. **21.** (c) Normal with mean $1/\theta$ and variance $1/[n\theta^3]$.

Chapter 9

Section 9.1

1. (a) $\pi(\beta|\delta) = e^{-\beta}$; (b) e^{-1}.
3. (a) $\pi(0) = 1$, $\pi(0.1) = 0.3941$, $\pi(0.2) = 0.1558$, $\pi(0.3) = 0.3996$, $\pi(0.4) = 0.7505$, $\pi(0.5) = 0.9423$, $\pi(0.6) = 0.9935$, $\pi(0.7) = 0.9998$, $\pi(0.8) = 1.0000$, $\pi(0.9) = 1.0000$, $\pi(1) = 1.0000$; (b) 0.1558. **5.** (a) Simple; (b) Composite; (c) Composite; (d) Composite. **9.** $T = \mu_0 - \overline{X}_n$. **11.** (a) $c_1 < 0$, $c_2 = 6$; (b) 0.0994. **13.** 3 **15.** $1 - x$, if $0 \le x \le 1$; 0, if $x > 1$. **19.** $(-\infty, \overline{x}_n + \sigma' n^{-1/2} T_{n-1}^{-1}(1-\alpha_0))$.

Section 9.2

1. Reject H_0 if $X = 1$; don't reject H_0 if $X = 0$. **3.** (b) 1. **5.** (a) Reject H_0 when $\overline{X}_n > 5 - 1.645n^{-1/2}$; (b) $\alpha(\delta) = 0.0877$. **7.** (b) $c = 31.02$. **9.** $\beta(\delta) = \left(\frac{1}{2}\right)^n$. **11.** (a) 0.6170; (b) 0.3173; (c) 0.0455; (d) 0.0027. **13.** (a) Reject H_0 if $\exp(-T/2)/4 < 4/(2+T)^3$; (b) Do not reject H_0; (d) Reject H_0 if $T > 13.28$; (e) Do not reject H_0.

Section 9.3

7. The power function is 0.05 for every value of θ. **9.** $c = 36.62$. **13.** (a) Reject H_0 if $\overline{X}_n \le 9.359$; (b) 0.7636; (c) 0.9995.

Section 9.4

1. $c_1 = \mu_0 - 1.645n^{-1/2}$ and $c_2 = \mu_0 + 1.645n^{-1/2}$. **3.** $n = 11$. **5.** $c_1 = -0.424$ and $c_2 = 0.531$. **11.** $c_1 = \mu_0 - 1.645n^{-1/2}$ and $c_2 = \mu_0 + 1.645n^{-1/2}$.

Section 9.5

1. (a) Don't reject H_0; (b) 0.0591. **3.** $U = -1.809$; do not reject the claim. **5.** Don't reject H_0. **9.** Since $\frac{S_n^2}{4} < 16.92$, don't reject H_0. **13.** $U = \frac{26}{3}$; the corresponding tail area is very small. **15.** $U = \frac{13}{3}$; the corresponding tail area is very small.

Section 9.6

1. Don't reject H_0. **3.** $c_1 = -1.782$ and $c_2 = 1.782$; H_0 will not be rejected. **5.** Since $U = -1.672$, reject H_0. **7.** $-0.320 < \mu_1 - \mu_2 < 0.008$. **11.** (a) Do not reject H_0; (b) Do not reject H_0.

Section 9.7

1. Reject the null hypothesis. 3. $c = 1.228$. 5. 1. 7. (a) $\hat{\sigma}_1^2 = 7.625$ and $\hat{\sigma}_2^2 = 3.96$; (b) Don't reject H_0. **9.** $c_1 = 0.321$ and $c_2 = 3.77$. **11.** $0.265V < r < 3.12V$. **15.** 0.8971. **19.** (a) 0.0503; (b) 0.0498.

Section 9.8

1. $X > 50.653$. **3.** Decide that failure was caused by a major defect if $\sum_{i=1}^{n} X_i > \frac{4n+\log(0.64)}{\log(7/3)}$. **11.** (a) For the first choice, $w_0 = w'$, $w_1 = w''$, $d_0 = d'$, $d_1 = d''$, $\Omega_0 = \Omega'$, and $\Omega_1 = \Omega''$. Switch them all for the other case.

Section 9.9

1. (a) $c = 1.96$. **3.** 0.0013. **5.** (a) 1.681, 0.3021, 0.25; (b) 0.0464, 0.00126, 3×10^{-138}.

Section 9.10

1. Reject H_0 if $X \geq 2$, $\alpha(\delta) = 0.5$, $\beta(\delta) = 0.1563$. **3.** Reject H_0 for $X \leq 6$. **5.** Reject H_0 for $X > 1 - \alpha^{1/2}$; $\beta(\delta) = (1 - \alpha^{1/2})^2$. **7.** Reject H_0 for $X \leq \frac{1}{2}[(1.4)^{1/2} - 1]$. **9.** Reject H_0 for $X \leq 0.01$ or $X \geq 1$; power is 0.6627. **11.** 0.0093. **17.** (a) 1; (b) $\frac{1}{\alpha}$. **23.** (a) Reject H_0 if the measurement is at least $5 + 0.1 \times \text{variance} \times \log(w_0\xi_0/[w_1\xi_1])$.

Chapter 10

Section 10.1

7. $Q = 11.5$; reject the hypothesis. **9.** (a) $Q = 5.4$ and corresponding tail area is 0.25; (b) $Q = 8.8$ and corresponding tail area is between 0.4 and 0.5.

Section 10.2

1. The results will depend on how one divides the real line into intervals, but the p-values for part (b) should be noticeably larger than the p-values for part (a). **3.** (a) $\hat{\theta}_1 = \frac{2N_1+N_4+N_5}{2n}$ and $\hat{\theta}_2 = \frac{2N_2+N_4+N_6}{2n}$. (b) $Q = 4.37$ and corresponding tail area is 0.226. **5.** $\hat{\theta} = 1.5$ and $Q = 7.56$; corresponding tail area lies between 0.1 and 0.2.

Section 10.3

1. $Q = 21.5$; corresponding tail area is 2.2×10^{-5}. **5.** $Q = 8.6$; corresponding tail area lies between 0.025 and 0.05.

Section 10.4

1. $Q = 18.8$; corresponding tail area is 8.5×10^{-4}. **3.** $Q = 18.9$; corresponding tail area is between 0.1 and 0.05. **5.** Correct value of Q is 7.2, for which the corresponding tail area is less than 0.05.

Section 10.5

7. (b)

	Proportion helped	
	Older subjects	Younger subjects
Treatment I	0.433	0.700
Treatment II	0.400	0.667

(c)

	Proportion helped
	All subjects
Treatment I	0.500
Treatment II	0.600

Section 10.6

3. $D_n^* = 0.25$; corresponding tail area is 0.11. **5.** $D_n^* = 0.15$; corresponding tail area is 0.63. **7.** $D_n^* = 0.065$; corresponding tail area is approximately 0.98. **9.** $D_{mn} = 0.27$; corresponding tail area is 0.39. **11.** $D_{mn} = 0.50$; corresponding tail area is 0.008.

Section 10.7

1. (a) 22.17; (b) 20.57, 22.02, 22.00, 22.00; (c) 22.10; (d) 22.00. **3.** 0.575. **5.** M.S.E. $(\overline{X}_n) = 0.025$ and M.S.E. $(\tilde{X}_n) = 0.028$. **13.** 1. **17.** Normal, with mean equal to the IQR of the distribution $(\theta_{3/4} - \theta_{1/4})$ and variance $[4nf(\theta_{1/4})^2]^{-1}$.

Section 10.8

3. $U = 3.447$; corresponding (two-sided) tail area is 0.003. **5.** $D_{mn} = 0.5333$; corresponding tail area is 0.010.

Section 10.9

1. (141, 175). **3.** Any level greater than 0.005, the smallest probability given in the table in this book. **5.** Do not reject the hypothesis. **9.** $|a| > \frac{1}{2}(6.635n)^{1/2}$. **15.** Normal, with mean $\left(\frac{1}{2}\right)^{1/\theta}$ and variance $\frac{1}{n\theta^2 4^{1/\theta}}$. **17.** (a) $0.031 < \alpha < 0.994$. (b) $\sigma < 0.447$ or $\sigma > 2.237$. **19.** Uniform on the interval $[y_1, y_3]$.

Chapter 11

Section 11.1

5. $y = -1.670 + 1.064x$. **7.** (a) $y = 40.893 + 0.548x$; (b) $y = 38.483 + 3.440x - 0.643x^2$. **9.** $y = 3.7148 + 1.1013x_1 + 1.8517x_2$. **11.** The sum of the squares of the deviations of the observed values from the fitted curve is smaller in Exercise 10.

Section 11.2

7. (a) −0.7861, 0.6850, 0.9377; (b) $0.2505\sigma^2$, $0.0277\sigma^2$; (c) −0.775 **9.** $c_1 = 3\overline{x}_n = 6.99$. **11.** $x = \overline{x}_n = 2.33$. **13.** −0.891. **15.** $c_1 = -\overline{x}_n = -2.25$. **17.** $x = \overline{x}_n = 2.25$.

Section 11.3

1. Since $U_0 = -6.695$, reject H_0. **3.** Since $U_1 = -6.894$, reject H_0. **5.** Since $|U_{01}| = 0.664$, don't reject H_0. **9.** Since $U^2 = 24.48$, reject H_0. **11.** $0.246 < \beta_2 < 0.624$. **13.** $0.284 < y < 0.880$. **17.** $10(\beta_1 - 0.147)^2 + 10.16(\beta_2 - 0.435)^2 + 8.4(\beta_1 - 0.147)(\beta_2 - 0.435) < 0.503$. **19.** $C = 1/(n-2)$. **25.** (a) $\beta_0 + \beta_1 x_i \pm T_{n-2}^{-1}(1 - \alpha_0/4)\sigma' \left[\frac{1}{n} + \frac{(x_i - \overline{x}_n)^2}{s_x^2}\right]^{1/2}$; (b) $\alpha(x) = \frac{x - x_1}{x_0 - x_1}$.

Section 11.4

5. (a) $12.21(\beta_1 - 0.4352)$ has the t distribution with eight degrees of freedom;
(b) $11.25(\beta_0 + \beta_1 - 0.5824)$ has the t distribution with eight degrees of freedom.

Section 11.5

5. $\hat{\beta} = 5.126, \hat{\sigma}^2 = 16.994$, and $\text{Var}(\hat{\beta}) = 0.0150\sigma^2$. **7.** $\hat{\beta}_0 = -0.744, \hat{\beta}_1 = 0.616, \hat{\beta}_2 = 0.013, \hat{\sigma}^2 = 0.937$. **9.** $U_3 = 0.095$; corresponding tail area is greater than 0.90. **11.** $R^2 = 0.644$. **13.** $\text{Var}(\hat{\beta}_0) = 222.7\sigma^2$, $\text{Var}(\hat{\beta}_1) = 0.1355\sigma^2$, $\text{Var}(\hat{\beta}_2) = 0.0582\sigma^2$, $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 4.832\sigma^2$, $\text{Cov}(\hat{\beta}_0, \hat{\beta}_2) = -3.598\sigma^2$, $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -0.0792\sigma^2$. **15.** $U_2 = 4.319$; corresponding tail area is less than 0.01. **21.** The value of the F statistic with two and seven degrees of freedom is 1.615; corresponding tail area is greater than 0.05. **25.** 87. **29.** 0.893.

Section 11.6

5. $U^2 = 13.09$; corresponding tail area is less than 0.025.

Section 11.7

5. $\mu = 3.25, \alpha_1 = -2, \alpha_2 = 3, \alpha_3 = -1, \beta_1 = 1.75, \beta_2 = -2.25, \beta_3 = -1.25, \beta_4 = 1.75$. **13.** $\hat{\sigma}^2 = 1.9647$. **15.** $U_B^2 = 4.664$; corresponding tail area is between 0.05 and 0.025.

Section 11.8

3. (a) $\mu = 9, \alpha_1 = -3, \alpha_2 = 3, \beta_1 = -1.5, \beta_2 = 1.5, \gamma_{11} = \gamma_{22} = \frac{1}{2}, \gamma_{12} = \gamma_{21} = -\frac{1}{2}$; (b) $\mu = 5, \alpha_1 = -\frac{1}{2}, \alpha_2 = \frac{1}{2}, \beta_1 = -\frac{3}{2}$, $\beta_2 = \frac{3}{2}, \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0$; (c) $\mu = 3\frac{1}{4}, \alpha_1 = -2, \alpha_2 = 3, \alpha_3 = -1, \beta_1 = 1\frac{3}{4}, \beta_2 = -2\frac{1}{4}, \beta_3 = -1\frac{1}{4}, \beta_4 = 1\frac{3}{4}, \gamma_{ij} = 0$ for all values of i and j; (d) $\mu = 5, \alpha_1 = -2\frac{1}{2}, \alpha_2 = 0, \alpha_3 = 2\frac{1}{2}, \beta_1 = -3, \beta_2 = -1, \beta_3 = 1, \beta_4 = 3, \gamma_{11} = 1\frac{1}{2}, \gamma_{12} = \frac{1}{2}, \gamma_{13} = -\frac{1}{2}$, $\gamma_{14} = -1\frac{1}{2}, \gamma_{21} = \gamma_{22} = \gamma_{23} = \gamma_{24} = 0, \gamma_{31} = -1\frac{1}{2}, \gamma_{32} = -\frac{1}{2}, \gamma_{33} = \frac{1}{2}, \gamma_{34} = 1\frac{1}{2}$. **11.** $U_{AB}^2 = 0.7047$; corresponding tail area is much larger than 0.05. **13.** $U_B^2 = 9.0657$; corresponding tail area is less than 0.025. **15.** The value of the appropriate statistic having the t distribution with 12 degrees of freedom is 2.8673; the corresponding tail area is between 0.01 and 0.005. **19.** $\alpha_0 + (1 - \alpha_0)\beta_0$.

Section 11.9

1. (a) (0.01996, 0.02129); (b) Reject the null hypothesis; (c) (25.35, 26.16). **3.** $E(T) = \frac{\rho\sigma_2}{\sigma_1}$; $\text{Var}(T) = \frac{(1-\rho^2)\sigma_2^2}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}$.

7. $\beta_2 = \dfrac{\sum_{i=1}^n (y_i'^2 - x_i'^2) \pm \left\{\left[\sum_{i=1}^n \left(y_i'^2 - x_i'^2\right)\right]^2 + 4\left(\sum_{i=1}^n x_i' y_i'\right)^2\right\}^{1/2}}{2\sum_{i=1}^n x_i' y_i'}$, $\beta_1 = \overline{y}_n - \beta_2 \overline{x}_n$, where $x_i' = x_i - \overline{x}_n$ and $y_i' = y_i - \overline{y}_n$. Either the plus sign or the minus sign in β_2 should be used, depending on whether the optimal line has a positive or a negative slope. **9.** $\frac{1}{n}\sum_{i=1}^k n_i \left[v_i^2 + (\overline{x}_{i+} - \overline{x}_{++})^2\right]$.

11. $\frac{1}{IJ(K-1)}\sum_{i,j,k}(Y_{ijk} - \overline{Y}_{ij+})^2$. **13.** Let $U = \frac{IJ(K-1)(S_A^2 + S_B^2 + S_{AB}^2)}{(IJ-1)S_{\text{Resid}}^2}$. Reject H_0 if $U \geq c$. Under H_0, U has an F distribution with $IJ - 1$ and $IJ(K-1)$ degrees of freedom. **15.** $\hat{\theta}_1 = \frac{1}{4}(Y_1 + Y_2) + \frac{1}{2}Y_3$, $\hat{\theta}_2 = \frac{1}{4}(Y_1 + Y_2) - \frac{1}{2}Y_3$, $\hat{\sigma}^2 = \frac{1}{3}[(Y_1 - \hat{\theta}_1 - \hat{\theta}_2)^2 + (Y_2 - \hat{\theta}_1 - \hat{\theta}_2)^2 + (Y_3 - \hat{\theta}_1 + \hat{\theta}_2)^2]$, where $Y_1 = W_1$, $Y_2 = W_2 - 5$, $Y_3 = \frac{1}{2}W_3$; $(\hat{\theta}_1, \hat{\theta}_2)$ and $\hat{\sigma}^2$ are independent; $(\hat{\theta}_1, \hat{\theta}_2)$ has a bivariate normal distribution with mean vector (θ_1, θ_2) and covariance matrix $\begin{bmatrix} \frac{3}{8} & -\frac{1}{8} \\ \frac{1}{8} & \frac{3}{8} \end{bmatrix}\sigma^2$; $\frac{3\hat{\sigma}^2}{\sigma^2}$ has the χ^2 distribution with one degree of freedom. **17.** $\text{Var}(\epsilon_i) = \left[1 - \frac{1}{n} - \frac{(x_i - \overline{x}_n)^2}{\sum_{j=1}^n (x_j - \overline{x}_n)^2}\right]\sigma^2$. **19.** $\mu = \overline{\theta} + \overline{\psi}$; $\alpha_i = \theta_i - \overline{\theta}$; and $\beta_j = \psi_j - \overline{\psi}$, where $\overline{\theta} = \frac{\sum_{i=1}^I v_i\theta_i}{v_+}$ and $\overline{\psi} = \frac{\sum_{j=1}^J w_j\psi_j}{w_+}$.

23.

$$\mu = \overline{\theta}_{+++},$$
$$\alpha_i^A = \overline{\theta}_{i++} - \overline{\theta}_{+++},$$
$$\alpha_j^B = \overline{\theta}_{+j+} - \overline{\theta}_{+++},$$
$$\alpha_k^C = \overline{\theta}_{++k} - \overline{\theta}_{+++},$$
$$\beta_{ij}^{AB} = \overline{\theta}_{ij+} - \overline{\theta}_{i++} - \overline{\theta}_{+j+} + \overline{\theta}_{+++},$$
$$\beta_{ik}^{AC} = \overline{\theta}_{i+k} - \overline{\theta}_{i++} - \overline{\theta}_{++k} + \overline{\theta}_{+++},$$
$$\beta_{jk}^{BC} = \overline{\theta}_{+jk} - \overline{\theta}_{+j+} - \overline{\theta}_{++k} + \overline{\theta}_{+++},$$
$$\gamma_{ijk} = \theta_{ijk} - \overline{\theta}_{ij+} - \overline{\theta}_{i+k} - \overline{\theta}_{+jk} + \overline{\theta}_{i++} + \overline{\theta}_{+j+} + \overline{\theta}_{++k} - \overline{\theta}_{+++}.$$

Chapter 12

Note: Answers to exercises that involve simulation are themselves only simulation approximations. Your answers will be different.

Section 12.1

5. (c) $f(x, y) = 0.4^3 x \exp(-0.4[x + y])$ for $x, y > 0$, $\int_0^\infty \int_x^\infty 0.4^3 x \exp(-0.4[x + y])dydx$.

Section 12.2

5. (b) The $k = 2$ trimmed mean probably has the smallest M.S.E. **9.** 0.2599. **11.** $(\lambda_{x1}\alpha_{x1}/\beta_{x1})^{1/2}(\mu_x - \mu_{x1})$ has the t distribution with $2\alpha_{x1}$ degrees of freedom, and similarly for μ_y. **15.** (a) $r - [\log \psi(1)]/u$.

Section 12.3

1. (a) Approximation = 0.0475, sim. std. err. = 0.0018; (b) $v = 484$. **11.** χ^2 distribution with $n - p$ degrees of freedom divided by S^2_{Resid}.

Section 12.4

7. (a) $Z = 0.8343$, sim. std. err. = 0.00372; (b) $Z' = 0.8386$, sim. std. err. = 0.00003. **17.** Look at Exercises 3, 4, 6, and 10.

Section 12.5

5. Approximation = 0.2542, sim. std. err. = 4.71×10^{-4}. **7.** 826.8, 843.3, 783.3. **9.** Means: −0.965, 0.02059, 1.199×10^{-5}; std. devs.: 2.448×10^{-2}, 1.207×10^{-4}, 8.381×10^{-6}. **11.** 0.33, 0.29, 0.30, 0.31, 0.34, 0.30, 0.62, 0.51, 0.98, 0.83. **13.** (b) Both α_0 and β_0 must be the same in both priors. In addition, $b_0 = \beta_0/\lambda_0$ and $a_0 = \alpha_0$; (d) Approximation = (154.67, 215.79), sim. std. err. of endpoints = $10.8v^{-1/2}$ (based on 10 Markov chains of length v each). **15.** (a) Conditional on everything else X_{n+i} has the d.f. $F(x) = [1 - e^{-\theta x}]/[1 - e^{-\theta c}]$, for $0 < x \le c$; (b) Conditional on everything else X_{n+i} has the d.f. $F(x) = 1 - e^{-\theta(x-c)}$, for $x \ge c$.

Section 12.6

3. $\sum_{i=(n+1)/2}^{n} \binom{n}{i} (\ell/n)^n (1 - \ell/n)^{n-i}$, where ℓ is the number of observations in the original sample that equal the smallest value. **5.** (a) −1.684; (b) About 50,000. **7.** (a) 0.107; (b) 1.763; (c) 0.0183. **9.** (a) 4.868×10^{-4}; (b) −0.0023; (c) 2.423×10^{-5} and 6.920×10^{-4}. **11.** (b) −0.2694 and 0.5458.

Section 12.7

5. (b) Approximation $= 0.581$, sim. std. err. $= 0.0156$; (c) 16,200. **7.** (a) 0.9 quantiles around 4.0, 0.95 quantiles around 5.2, 0.99 quantiles around 8; (b) The differences are on the same order of magnitude as Monte Carlo variability; (c) 0.123. **9.** (a) $\exp\left(-\beta\phi_0 - \frac{u_0(\psi-\psi_0)^2}{2} - \sum_{i=1}^{p}\tau_i\left[\beta + \frac{n_i(\mu_i-\overline{y}_i)^2+w_i+\lambda_0(\mu_i-\psi)^2}{2}\right]\right)$ $\times\beta^{p\alpha_0+\epsilon_0-1}\prod_{i=1}^{p}\tau_i^{\alpha_0+[n_i+1]/2-1}$; (b) β has a gamma distribution with parameters $p\alpha_0+\epsilon_0$ and $\phi_0+\sum_{i=1}^{p}\tau_i$; (c) Very close to the values in Table 12.6. **11.** (c) The proportions are rather close to the nominal values.

This page intentionally left blank

References

Allison, D. B., Heshka, S., Sepulveda, D., and Heymsfield, S. B. (1993). Counting calories—Caveat emptor. *Journal of the American Medical Association*, 270: 1454–1456.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton: Princeton University Press.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley and Sons.

Berry, D. A., and Geisser, S. (1986). Inference in cases of disputed paternity. In M. H. DeGroot, S. E. Fienberg, and J. B. Kadane (eds.), *Statistics and the Law* (pp. 353–382). New York: John Wiley and Sons.

Bickel, P. J., and Doksum, K. A. (2000). *Mathematical Statistics: Basic Ideas and Selected Topics* (Vol. 1, 2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81: 637–654.

Bortkiewicz, L. von, (1898). *Das Gesetz der Kleinen Zahlen*. Leipzig: Teubner.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Box, G. E. P., and Müller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 29: 610–611.

Broemeling, L. (1984). *Bayesian Analysis of Linear Models*. New York: Marcel Dekker, Inc.

Brunk, H. D. (1975). *An Introduction to Mathematical Statistics* (3rd ed.). Lexington, MA: Xerox College Publishing.

Casella, G., and Berger, R. L., (2002) *Statistical Inference* (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Chase, M. A., and Dummer, G. M. (1992). The role of sports as a social status determinant for children. *Research Quarterly for Exercise and Sport*, 63: 418–424.

Cook, R. D., and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

——— (1994). *An Introduction to Regression Graphics*. New York: John Wiley and Sons.

——— (1999). *Applied Regression Including Computing and Graphics*. New York: John Wiley and Sons.

Cowles, M. K., and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91: 883–904.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Cullen, C. G. (1972). *Matrices and Linear Transformations* (2nd ed.). Reading, MA: Addison-Wesley.

Darwin, C. (1876). *The Effects of Cross and Self-Fertilization in the Vegetable Kingdom*. London: John Murray.

David, F. N. (1988). *Games, Gods, and Gambling*. New York: Dover Publications.

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. New York: Cambridge University Press.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

Devore, J. L. (1999). *Probability and Statistics for Engineering and the Sciences* (5th ed.). Monterey, CA: Brooks/Cole.

Devroye, L. (1986). *Non-Uniform Random Variate Generation* New York: Springer.

Draper, N. R., and Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). New York: John Wiley and Sons.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7: 1–26.

Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (Vol. 1, 3rd ed.). New York: John Wiley and Sons.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.

Finkelstein, M. O., and Levin, B. (1990). *Statistics for Lawyers*. New York: Springer-Verlag.

Forbes, J. D. (1857). Further experiments and remarks on the measurement of heights by the boiling point of water. *Transactions of the Royal Society of Edinburgh*, 21: 135–143.

Fraser, D. A. S. (1976). *Probability and Statistics*. Boston: Duxbury Press.

Frey, M., and Edwards, M. (1997). Surveying arsenic occurrence. *Journal of the American Water Works Association*, 89: 105–117.

Friedland, L. R., Joffe, M., Wiley, J. F., Schapire, A., and Moore, D. F., (1992). Effect of educational program on compliance with glove use in a pediatric emergency department. *American Journal of Diseases of Childhood*, 146: 1355–1358.

Frisby, J. P., and Clatworthy, J. L. (1975) Learning to see complex random-dot stereograms, *Perception*, 4: 173–178.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85: 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6: 721–741.

Graybill, F. A., and Iyer, H. K. (1994). *Regression Analysis: Concepts and Applications*. Pacific Grove, CA: Brooks/Cole.

Grunbaum, B. W., Crim, M., Selvin, S., Pace, N., and Black, D. M. (1978). Frequency distribution and discrimination probability of twelve protein genetic variants in human blood as functions of race, sex, and age. *Journal of Forensic Sciences*, 23: 577–587.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley and Sons.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57: 97–109.

Heavenrich, R. M., and Hellman, K. H. (1999) Light duty automotive technology and fuel economy trends through 1999, *U.S. Environmental Protection Agency* (EPA420-R-99-018).

Hoel, P. G., Port, S., and Stone, C. L. (1971). *Introduction to Probability Theory*. Boston: Houghton-Mifflin.

Hogg, R. V., and Tanis, E. A. (1997). *Probability and Statistical Inference* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35: 73–101.

——— (1977). *Robust Statistical Procedures*. Philadelphia: Society for Industrial and Applied Mathematics.

——— (1981). *Robust Statistics*. New York: John Wiley and Sons.

Kempthorne, O., and Folks, L. (1971). *Probability, Statistics, and Data Analysis*. Ames, IA: Iowa State University Press.

Kennedy, W. J., Jr., and Gentle, J. E. (1980). *Statistical Computing*. New York: Marcel Dekker.

Kronmal, R. A., and Peterson, A.V., Jr. (1979). On the alias method for generating random variables from a discrete distribution. *The American Statistician*, 33: 214–218.

Larsen, R. J., and Marx, M. L. (2001). *An Introduction to Mathematical Statistics and Its Applications* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Larson, H. J. (1974). *Introduction to Probability Theory and Statistical Inference* (2nd ed.). New York: John Wiley and Sons.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: John Wiley and Sons.

Lehmann, E. L. (1958). Significance level and power. *Annals of Mathematical Statistics*, 29: 1167–1176.

——— (1997). *Testing Statistical Hypotheses* (2nd ed.). New York: Springer-Verlag.

Lehmann, E. L., and Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). New York: Springer-Verlag.

Lieblein, J., and Zelen, M. (1956). Statistical investigation of the fatigue life of deep groove ball bearings. *Journal of Research of the National Bureau of Standards*, 57: 273–316.

Lindgren, B. W. (1976). *Statistical Theory* (3rd ed.). New York: Macmillan.

Lockwood, J. R., Schervish, M. J., Gurian, P., and Small, M. J. (2001). Characterization of arsenic occurrence in source waters of U.S. community water systems. *Journal of the American Statistical Association*, 96(456): 1184–1193.

Lorenzen, T. J. (1980). Determining statistical characteristics of a vehicle emissions audit procedure. *Technometrics*, 22: 483–493.

Lubischew, A. A. (1962). On the use of discriminant functions in taxonomy. *Biometrics*, 18: 455–477.

Lyle, R. M., Melby, C. L., Hyner, G. C., Edmondson, J. W., Miller, J. Z., and Weinberger, M. H. (1987). Blood pressure and metabolic effects of calcium supplementation in normotensive white and black men. *Journal of the American Medical Association*, 257: 1772–1776.

Manly, B. F. J. (1986). *Multivariate Statistical Methods: A Primer*. London: Chapman and Hall.

Markowitz, H. (1987). *Mean-Variance Analysis in Portfolio Choice and Capital Markets*. Oxford: Basil Blackwell.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21: 1087–1091.

Meyer, P. L. (1970). *Introductory Probability and Statistical Applications* (2nd ed.). Reading, MA: Addison-Wesley.

Miller, I., and Miller, M. (1999). *John E. Freund's Mathematical Statistics* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the Theory of Statistics* (3rd ed.). New York: McGraw-Hill.

Moore, D. S., and McCabe, G. P. (1999). *Introduction to the Practice of Statistics* (3rd ed.). New York: W. H. Freeman.

Morrison, D. F. (1990). *Multivariate Statistical Methods* (3rd ed.). New York: McGraw-Hill.

Nocedal, J., and Wright, S. (2006). *Numerical Optimization*. New York: Springer.

Olkin, I., Gleser, L. J., and Derman, C. (1980). *Probability Models and Applications*. New York: Macmillan.

Ore, O. (1960). Pascal and the invention of probability theory. *American Mathematical Monthly*, 67: 409–419.

Prien, R. F., Kupfer, D. J., Mansky, P. A., Small, J. G., Tuason, V. B., Voss, C. B., and Johnson, W. E. (1984). Drug therapy in the prevention of recurrences in unipolar and bipolar affective disorders. *Archives of General Psychiatry*, 41: 1096–1104.

Quetelet, A. (1846). *Lettres à S.A.R. le Duc Régnant de Saxe-Cobourg et Gotha, sur la théorie des probabilités, appliquée aux sciences morales et politiques*. Brussels: Hayez.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (2nd ed.). New York: John Wiley and Sons.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis* (2nd ed.). Belmont, CA: Duxbury Press.

Rohatgi, V. K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. New York: John Wiley and Sons.

Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley and Sons.

Rutherford, E., and Geiger, H. (1910). The probability variations in the distribution of α particles. *The London, Dublin, and Edinburgh Philosophical Magazine and Journal of Science, Series 6*, 20: No. 118, 698–704.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method*. New York: John Wiley and Sons.

Scheffé, H. (1959). *The Analysis of Variance*. New York: John Wiley and Sons.

Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer-Verlag.

Sharpe, R. H., and Van Middelem, C. H. (1955). Application of variance components to horticultural problems with special reference to a parathion residue study. *Proceedings of the American Society for Horticultural Science*, 66: 415–420.

Simpson, J., Olsen, A., and Eden, J. C. (1975). A Bayesian analysis of a multiplicative treatment effect in weather modification. *Technometrics*, 17: 161–166.

Smith, H. L., Piland, N. F., and Fisher, N. (1992). A comparison of financial performance, organizational characteristics, and management strategy among rural and urban nursing facilities. *Journal of Rural Health*, 8: 27–40.

Sokal, R. R., and Rohlf, F. J. (1981). *Biometry* (2nd ed.). San Francisco: W. H. Freeman.

Stigler, S. M. (1986). *The History of Statistics*. Cambridge, MA: Belknap Press of Harvard University Press.

Student (1908). The probable error of a mean. *Biometrika*, 6: 1–25.

Thomson, A., and Randall-Maciver, R. (1905). *Ancient Races of the Thebaid*. Oxford: Oxford University Press.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, 22: 1701–1762.

Todhunter, I. (1865). *A History of the Mathematical Theory of Probability from the Time of Pascal to That of Laplace*. New York: G. E. Stechert (reprinted 1931).

Tubb, A., Parker, A. J., and Nickless, G. (1980). The analysis of Romano-British pottery by atomic absorption spectrophotometry. *Archaeometry*, 22: 153–171.

Twain, M. (1924). *Mark Twain's Autobiography* (Vol. 1). New York: Harper Brothers.

Wackerly, D. D., Mendenhall, W., and Scheaffer, R. L. (1996). *Mathematical Statistics with Applications* (7th ed.). Belmont, CA: Duxbury Press.

Walker, A. J. (1974). New fast method for generating discrete random numbers with arbitrary frequency distributions. *Electronics Letters*, 10: 127–128.

Weisberg, S. (1985). *Applied Linear Regression* (2nd ed.). New York: John Wiley and Sons.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29: 350–362.

——— (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 29: 28–35.

——— (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, 29: 330–336.

Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics*, 1: 80–83.

Winsor, C. P. (1947). Quotations: "Das Gesetz der Kleinen Zahlen," *Human Biology*, 19: 154–161.

Young, G. A. (1994). Bootstrap: More than a stab in the dark? (with discussion). *Statistical Science*, 9: 382–415.

This page intentionally left blank

INDEX

Discrete Distributions

	Bernoulli with parameter p	Binomial with parameters n and p
p.f.	$f(x) = p^x(1-p)^{1-x}$, for $x = 0, 1$	$f(x) = \binom{n}{x} p^x(1-p)^{n-x}$, for $x = 0, \ldots, n$
Mean	p	np
Variance	$p(1-p)$	$np(1-p)$
m.g.f.	$\psi(t) = pe^t + 1 - p$	$\psi(t) = (pe^t + 1 - p)^n$

	Uniform on the integers $a, \ldots, b$	Hypergeometric with parameters A, B, and n
p.f.	$f(x) = \frac{1}{b-a+1}$, for $x = a, \ldots, b$	$f(x) = \frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}}$, for $x = \max\{0, n-b\}, \ldots, \min\{n, A\}$
Mean	$\frac{b+a}{2}$	$\frac{nA}{A+B}$
Variance	$\frac{(b-a)(b-a+2)}{12}$	$\frac{nAB}{(A+B)^2}\frac{A+B-n}{A+B-1}$
m.g.f.	$\psi(t) = \frac{e^{(b+1)t} - e^{at}}{(e^t-1)(b-a+1)}$	Nothing simpler than $\psi(t) = \sum_x f(x)e^{tx}$

	Geometric with parameter p	Negative binomial with parameters r and p
p.f.	$f(x) = p(1-p)^x$, for $x = 0, 1, \ldots$	$f(x) = \binom{r+x-1}{x} p^r(1-p)^x$, for $x = 0, 1, \ldots$
Mean	$\frac{1-p}{p}$	$\frac{r(1-p)}{p}$
Variance	$\frac{1-p}{p^2}$	$\frac{r(1-p)}{p^2}$
m.g.f.	$\psi(t) = \frac{p}{1-(1-p)e^t}$, for $t < \log(1/[1-p])$	$\psi(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r$, for $t < \log(1/[1-p])$

	Poisson with mean λ	Multinomial with parameters n and $(p_1, \ldots, p_k)$
p.f.	$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$, for $x = 0, 1, \ldots$	$f(x_1, \ldots, x_k) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k}$, for $x_1 + \cdots + x_k = n$ and all $x_i \geq 0$
Mean	λ	$E(X_i) = np_i$, for $i = 1, \ldots, k$
Variance	λ	$\text{Var}(X_i) = np_i(1-p_i)$, $\text{Cov}(X_i, X_j) = -np_ip_j$, for $i, j = 1, \ldots, k$
m.g.f.	$\psi(t) = e^{\lambda(e^t-1)}$	Multivariate m.g.f. can be defined, but is not defined in this text.

Continuous Distributions

	Beta with parameters α and β	Uniform on the interval $[a, b]$
p.d.f.	$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, for $0 < x < 1$	$f(x) = \frac{1}{b-a}$, for $a < x < b$
Mean	$\frac{\alpha}{\alpha+\beta}$	$\frac{a+b}{2}$
Variance	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\frac{(b-a)^2}{12}$
m.g.f.	Not available in simple form	$\psi(t) = \frac{e^{-at}-e^{-bt}}{t(b-a)}$

	Exponential with parameter β	Gamma with parameters α and β
p.d.f.	$f(x) = \beta e^{-\beta x}$, for $x > 0$	$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, for $x > 0$
Mean	$\frac{1}{\beta}$	$\frac{\alpha}{\beta}$
Variance	$\frac{1}{\beta^2}$	$\frac{\alpha}{\beta^2}$
m.g.f.	$\psi(t) = \frac{\beta}{\beta-t}$, for $t < \beta$	$\psi(t) = \left(\frac{\beta}{\beta-t}\right)^\alpha$, for $t < \beta$

	Normal with mean μ and variance σ^2	Bivariate normal with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ
p.d.f.	$f(x) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	Formula is too large to print here. See Eq. (5.10.2) on page 338.
Mean	μ	$E(X_i) = \mu_i$, for $i = 1, 2$
Variance	σ^2	Covariance matrix: $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$
m.g.f.	$\psi(t) = \exp\left(\mu t + \frac{t^2\sigma^2}{2}\right)$	Bivariate m.g.f. can be defined, but is not defined in this text.

Continuous Distributions

	Lognormal with parameters μ and σ^2	F with m and n degrees of freedom
p.d.f.	$f(x) = \frac{1}{(2\pi)^{1/2}\sigma x} \exp\left(-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right)$, for $x > 0$	$f(x) = \frac{\Gamma\left[\frac{1}{2}(m+n)\right] m^{m/2} n^{n/2}}{\Gamma\left(\frac{1}{2}m\right)\Gamma\left(\frac{1}{2}n\right)} \cdot \frac{x^{(m/2)-1}}{(mx+n)^{(m+n)/2}}$, for $x > 0$
Mean	$e^{\mu+\sigma^2/2}$	$\frac{n}{n-2}$, if $n > 2$
Variance	$e^{2\mu+\sigma^2}[e^{\sigma^2} - 1]$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$, if $n > 4$
m.g.f.	Not finite for $t > 0$	Not finite for $t > 0$

	t with m degrees of freedom	χ^2 with m degrees of freedom
p.d.f.	$f(x) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{(m\pi)^{1/2}\Gamma\left(\frac{m}{2}\right)} \left(1 + \frac{x^2}{m}\right)^{-(m+1)/2}$	$f(x) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{(m/2)-1} e^{-x/2}$, for $x > 0$
Mean	0, if $m > 1$	m
Variance	$\frac{m}{m-2}$, if $m > 2$	$2m$
m.g.f.	Not finite for $t \neq 0$	$\psi(t) = (1 - 2t)^{-m/2}$, for $t < 1/2$

	Cauchy centered at μ	Pareto with parameters x_0 and α_0
p.d.f.	$f(x) = \frac{1}{\pi(1+[x-\mu]^2)}$	$f(x) = \frac{\alpha x_0^\alpha}{x^{\alpha+1}}$, for $x > x_0$
Mean	Does not exist	$\frac{\alpha x_0}{\alpha-1}$, if $\alpha > 1$
Variance	Does not exist	$\frac{\alpha x_0^2}{(\alpha-1)^2(\alpha-2)}$, if $\alpha > 2$
m.g.f.	Not finite for $t \neq 0$	Not finite for $t > 0$